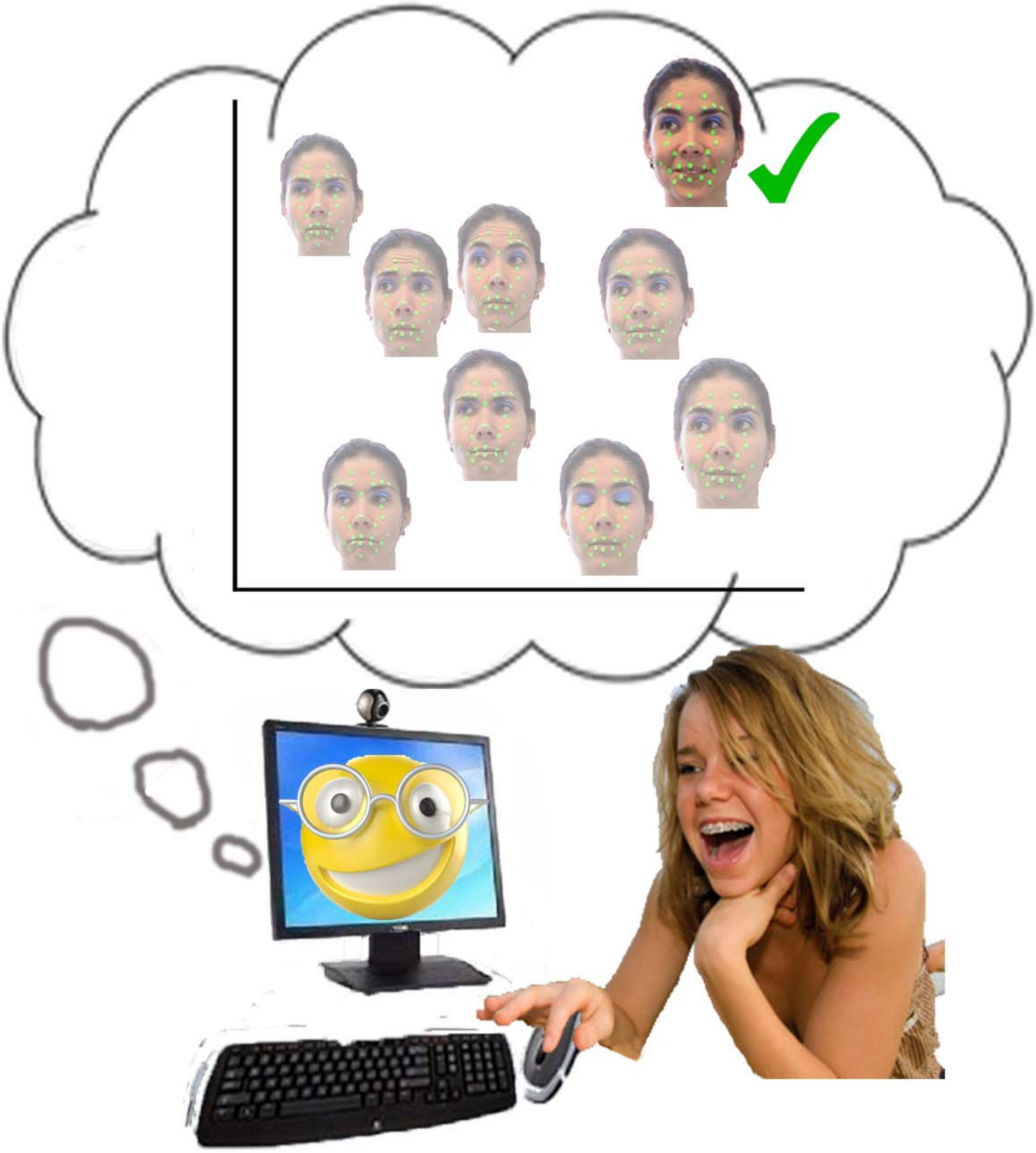
EXTRACTING EMOTIONS FROM FACE-TO-FACE COMMUNICATION



Pegah Takapoui
July 2009

Delft University of Technology

Extracting emotions from face-to-face communication

In partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Pegah Takapoui
Born in Kermanshah, Iran

July 2009

Man Machine Interaction group
Delft University of Technology
Faculty of Electrical Engineering, Mathematics, and Computer Science
Mekelweg 4
2628 CD Delft, The Netherlands
http://www.ewi.tudelft.nl

Extracting emotions from face-to-face communication

In our life we get more and more dependent on our computer and we have less time for face-to-face social activities with friends and families. In face-to-face communication our faces convey lots of emotions using facial expressions and lots of information is transmitted faster non-verbally than verbally, through the facial expressions. Using lots of modalities like facial expressions, speech and (hand and body) gestures make the communication between the humans multimodal. But the current communication style between human and computer is still dominated by keyboard and mouse and there is no room for emotions and facial expressions. In order to have a better understanding of human-human communication and to improve the human-computer interaction it is essential to identify and describe the different modalities of human-human communication including collection and annotation of multimodal data. Using the facial expressions, (hand and body) gestures, speech recognition and content awareness make the communication multimodal and enable the computer to adapt itself to the needs of the individual users.

In this report we study one of the important modalities, the facial expressions and we proposed an algorithm for tracking the facial expressions from face-to-face communication. To discover the relation between the different facial expressions and their meaning we needed data of the human face-to-face communication to analyze the facial expressions during the interaction between them. After some research we decided to make some recordings and build our own database. Our research problem is to localize facial expressions, to label them and research the communication impact. To facilitate the localization we put markers on the faces of test persons during our experiment. We asked 3 observers (annotators) to watch the recordings of the face-to-face communication and put labels on the segments that contain an emotion. Calculating the level of agreement between the annotators we compared their results with each other and finally we used these labeled segments in our model to extract the different features for each emotion. Using our model we followed the changes of the face during the facial expressions to collect facial features and used these facial features to define emotion clusters. Using these clusters we can start building an automatic emotion extraction tool from face-to-face communications. For defining enough features to recognize the emotions we need a much bigger database. During this project using our data, we extract some features for basic emotions and define clusters to recognize them, like recognizing 'sadness' from 'happiness' and 'anger' from 'surprise'. For making a big more emotion clusters and recognizing more emotions we need more annotated data.

Thesis Committee:

Chair:	Prof. Dr. Drs. L. J. M. Rothkrantz, Faculty EEMCS, TU Delft
Committee Member:	Ir. H. Geers, Faculty EEMCS, TU Delft
Committee Member:	Dr. Ir. P. Wiggers, Faculty EEMCS, TU Delft
Committee Member:	Ir. A. G. Chitu, Faculty EEMCS, TU Delft
Committee Member:	Dr. Ir. C.A.P.G van der Mast

Acknowledgements

First of all I would like to express my gratitude to my supervisor Prof. Drs. Dr. L.J.M. Rothkrantz, for giving me the opportunity to do this research assignment and guiding me through the process. Besides my daily supervisor I would like to thank Ir. A. G. Chitu for sharing his experience with me. I would also like to thank my fellow students and friends, with which I have experienced this period of my life, for all their input and companionship especially dear Anoop.

I also want to thank my parents, for their unfailing support from distance and having faith in me all the time. And finally especial thanks to the love of my life, my dear Shahin who stood by me and backed me up every day for the last 10 years. Thank you my dear!

TABLE OF CONTENTS

Acknowledgements 7
1. Introduction 15
1.1. Non verbal communication 16
1.2. Facial expressions 16
1.3. Problem overview 19
1.4. Research goals 20
2. Literature survey 23
2.1. Annotation tools 23
2.2. The MATE Annotation workbench 25
2.3. Microsoft Research Annotation System (MRAS) 27
2.4. SmartKom 28
2.5. Ruth 32
3. Data collection 35
3.1. The New Delft University of Technology Audio-Visual Speech Corpus 36
3.2. Anna's recordings 38
4. Annotation 43
4.1. Experiment 43
4.2. Data Analysis 45
4.3. The offered labels 48
4.4. Triggers 49
4.5. Duration and frequency 50
5. Model of tracking points 53
5.1. Prototype 54
5.2. Labeling and validation 56
6. Results of tracking 59
6.1. Tracking of selected points 60
6.1.1. Points on the eyebrows 61
6.1.2. Points around the mouth 63
6.2. Principal component analysis 64
6.3. Euclidean distance matrix 67
7. Models of human face-to-face communication 73
7.1. Basic face-to-face communication models 73
7.2. Causal model of interactions 76
7.3. Bayesian model of interaction 78
7.4. Validation of developed interaction models 79
8. Conclusions and future work 81
8.1. Concluding remarks 83
8.2. Future work 84
Bibliographies 85
Appendices 89
A: Annotation protocol 89
B: The annotated recordings 91
C: Recording protocol 103
D: Assigned labels by annotators 113

LIST OF FIGURES

Figure 1: Shahrdzad, the storyteller tries to postpones her execution by telling a story 15
Figure 2: Basic emotions. 18
Figure 3: An example of Benesh movement notation. 24
Figure 4: Different components of an annotation tool to detect aggression in trains. 25
Figure 5: Annotation tool DAT. 26
Figure 6: Adding an audio annotation to a video recording using MRAS. 27
Figure 7: User can use the 'Query Annotations' box to search for a annotation. 28
Figure 8: Delegation-oriented dialogue paradigm of Smartakus. 29
Figure 9: Multimodal input and output in Smartkom.. 30
Figure 10: Ruth, the talking head.. 32
Figure 11: The setup of the recording room. 37
Figure 12: Set up of Anna's recording.. 39
Figure 13: Typical appearance of facial expression: ‘astonishment’. 40
Figure 14: Annotating a recording using ELAN. 44
Figure 15: Number of occurrences of different emotional states done by first annotator. 52
Figure 16: Number of occurrences of different emotional states done by second annotator. 52
Figure 17: The original image and the detected green points on the face by processing. 55
Figure 18: The original image of the face and the detected blue blobs by processing. 55
Figure 19: Flowchart of labeling process. 56
Figure 20: Dependencies between points. 57
Figure 21: Following the movements on the face to find the 59
Figure 22: Point 4 and point 24. 60
Figure 23: Following point 4 in a recording containing a transition from 'disbelief' to 'grief'. 61
Figure 24: Movement of point4 during emotional state 'surprise'. 62
Figure 26: Distances between point 5 on apex frame and other frames. 62
Figure 25: Distances between point 4 on apex frame and other frames. 62
Figure 27: Movements of point 24 and 25 during 'happiness'. 63
Figure 28: Following of the movements of point 24 during 'happiness'. 63
Figure 29: Following of the movements of point 25 during 'happiness'. 64
Figure 30: First and second principal component at some frames before apex frame. 66
Figure 31: First and second principal component at apex frame. 66
Figure 32: First and second principal component at some frame after apex frame. 66
Figure 33: Labels from our table, rated using Whissell’s database. 70
Figure 34: S-R mechanism between mother and baby.. 74
Figure 35: Context sensitive S-R model. 74
Figure 36: Markov model of emotional states. 76
Figure 39: A text fragment with some emotional words.. 77
Figure 37: Emotional states. 77
Figure 38: Emotional words.. 77
Figure 40: Emotional words of one sentence and the average sum of them 78
Figure 41: Average sum of the emotional words and the Gaussian distribution of 2 words......... 78
Figure 42: At 1001th night sultan decides to spare the life of Shahrdzad 81
Figure 43: Template expressions. 90

LIST OF TABLES

Table 1: Actions and goals. 20
Table 2: Situation and reactions to elicit 'Amusement'. 38
Table 3: Different type of recordings and the reactions. 39
Table 4: A fragment from Anna's recordings. 40
Table 5: Results of calculating the inter-label agreements. 46
Table 6: Categorized labels. 47
Table 7: Results of recalculating the inter-label agreements. 48
Table 8: Template expressions. 48
Table 9: Statics of manually selected facial expressions by two annotators. 51
Table 10: Euclidean distances between different emotions from our recordings. 68
Table 11: Distribution of distances per emotion 69
Table 12: The emotion’s pleasantness level and arousal level. 69
Table 13: Euclidean distances between different emotions from Whissell dictionary. 71
Table 14: Example of facial expressions triggered by the keywords. 75

Chapter 1

1. Introduction

Our lifestyles nowadays have changed from 20 years ago. People are much busier and have no time for social interactions. Contacts with friends and family are less and less face-to-face each day and more and more with the help of computer and telephone. In the human computer interaction of these days there is no room for emotions. If we could communicate with computers the way we do in face-to-face communication lots of information could be conveyed through the context and emotion of the speaker.

When I was young my mother told me lots of stories. Historical tales, love stories, tragedies and comedies. There were lots of emotions in those stories and I sympathized with all the characters of the stories. My favorite story was the story about Shahrdzad the storyteller. On the night of her marriage with the bitter sultan who wants to execute her the following morning, she starts telling the sultan a story. The sultan is so intrigued by her storytelling that he spares her life each day so he can continue to enjoy her story. The story goes on for 1001 nights. I loved those stories and learned a lot about life and human emotions through them. But I wonder if the children of this time could experience the taste of having bedtime stories told by their parents.

Figure 1: Shahrdzad, the storyteller tries to postpones her execution by telling the sultan a story (Lang, 1898).

In our life we have less time for face-to-face social activities with friends and families and get more and more dependent on our computer. But the communication between us and computer is still dominated by keyboard and mouse. If we could at least improve the communication methods with computer so that human emotions were involved, maybe we could build a multimedia story teller for our children which is able to tell stories for our children and show the matching emotion to enhance the interaction and convey the feeling to child. Then we can expect the computer to adapt itself to the situation and the needs of children at that moment in the same way as it happens when we talk face-to-face. A lot of information is extracted from the context and emotion of the speaker (Pantic M., September 2003).

Almost all of our daily activities involve communication. We communicate interactively and in multimodal manner. Communication in general is a process of sending and receiving messages that enables human to share knowledge and skills and is composed of two dimensions, verbal and nonverbal (Pantic M., October 2000).

1.1. Non verbal communication

Nonverbal communication has been defined as communication without words and the facial expressions are the basic mode of nonverbal communication. Nonverbal communications also include eye contact, gesture and body language. According to the theory of Dr. Mehrabian during human communication verbal cues provide just 7 percent of the meaning of the message, vocal cues 38 percent and facial expressions 55 percent. This means that the receiver of a message can rely heavily on the facial expressions of the sender because his expressions are a better indicator of the meaning behind the message than his words (Mehrabian, June 1967). Of course *"7%-38%-55%"* rule should not be overly interpreted because we are using vocal aspect of communications most of the time unless a communicator is talking about his/her feelings or attitudes.

Speech may contain nonverbal elements such as emotion, speaking style, rhythm, intonation and stress. Besides, some facial expressions can even replace words, for example nodding the head can replace a verbal communication. Not only speaker but listener also uses facial expression during a conversation and gives non verbal feedback to speaker. The human facial expressions also help to understand the speech better and faster. In a noisy environment the lip reading and recognizing the facial expressions contribute to a better communication. The facial expressions also change according to the content of a message and the flow of a conversation.

1.2. Facial expressions

Emotions and emotional states have different meaning. The concept of emotions consists of feelings, which are subjective experience of an emotion. An emotional state is measured through various physiological changes in the body and especially in the face as a response to an emotion. The physical changes are not just limited with facial expressions but include also changes in the blood pressure or muscle tension. However, facial expressions remain the most important channel to express the emotional states.

Paul Ekman was one of the researchers who tried to find a relationship between facial expressions and emotions. According to his theory, emotions expressed with facial expressions act as social signals and help people to communicate (Wojdel A., 2005). Ekman distinguishes 6 basic emotions which can be easily distinguished from each other. These six basic emotions are: anger, disgust, fear, happiness, sadness, and surprise. These basic emotions are innate and have universal facial expressions across different cultures and act as basic building blocks of all emotions. Together with Friesen they gave precise description of facial features corresponding to each basic emotion, their blending and how they differ depending on the intensity (Ekman, 1975). Below the descriptions of these six basic emotions are given.

Anger: Anger can be provoked e.g. by frustration, physical threat, or feeling of being hurt by somebody. It varies in intensity from irritation to fury. Person experiencing anger can behave violently. Anger is expressed on the face with eyebrows lowered and drawn together, eyes opened and staring in one direction, lips hard pressed together or parted in square shape. This expression is mostly shown with eyebrows which are drowning together. The eyes are opened and staring in one direction (see Figure 1a).

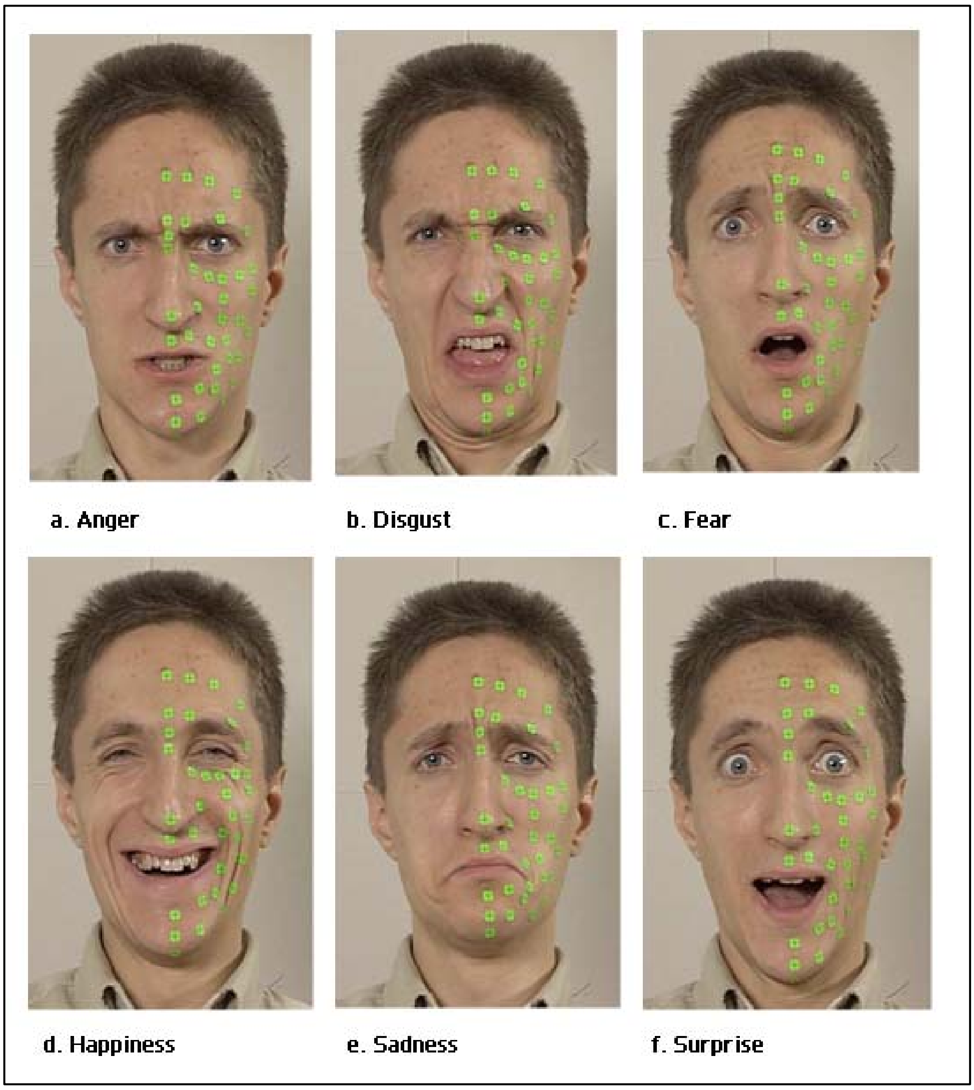
Disgust: Disgust involves a feeling of aversion to taste, smell, touch, appearance, or some action. Response for mild disgust – dislike – is a wish to turn away from the disgusting object, while extreme disgust can be even a reason for vomiting. This emotion manifests itself with raising the upper lip, wrinkling the nose, and lowering the eyebrows (see Figure 1b).

Fear: Fear occurs when a person is expecting some event which can physically or psychologically harm his/her. It ranges from apprehension to terror. In the intensive form, it is the most traumatic of all emotions. Fear is characterized with eyebrows raised and drown together and the lips stretched back. Eyes are usually opened with lower lid tensed (see Figure 1c).

Happiness: Happiness is the most positive emotion. People often experience happiness together with states of excitement, pleasure, or relief. Happiness is primarily expressed with mouth: corners of the lips are raised, and nasolabial folds are deepened. In extreme happiness, eyes are narrower with crows-feet wrinkles appearing around their outer corners (see Figure 1d).

Sadness: This emotion is not as lots of people think the opposite of the happiness. It is a feeling of suffering caused by loss, disappointment, or hopelessness. It may last for a very long time – hours, or even days. It varies from a feeling of gloom to deep mourning. Sad person expresses the emotion by the fact that the inner corners of eyebrows are raised and drawn together, lower eyelids are little bit raised, and the corners of the lips are pulled downwards (see Figure1e).

Surprise: Surprise is evoked by unexpected or misexpected event. It is a short term expression (when a person has time to think about surprising event he/she is not surprised anymore). It manifests itself with raised eyebrows, eyes wide open and jaw dropped causing parting of the lips (see Figure 1f).



a. Anger

b. Disgust

c. Fear

d. Happiness

e. Sadness

f. Surprise

Figure 2: Basic emotions.

The six basic emotions exist in all different cultures and they are innate. But except these six basic emotions there are lots of other emotions that are expressed by people everywhere every day. Those emotions are culture dependent, they are learnt during life and they are built up from the basic emotions. They form the families of the basic emotions (Ekman, 1975). Each member of emotion family shares certain characteristics. For example they share the same physiological activities or the antecedent events.

These shared characteristics of the family distinguish one family from another. For example anger is specified with more than 60 different facial expressions and they are all from the family anger. They share same features, and they differ from all other families. In the case of anger, the eyebrows are lowered and drawn together, the upper eyelid is raised and the tightened muscles in the lip make these emotions al member of the same family. Each family of emotion has his own characteristic.

Members of different cultures have the common facial expressions experiencing the same emotion when it comes to six basic emotions. There are few studies which have compared the facial expressions actually produced by member of different cultures in comparable situations. Ekman studied (Ekman P., 1972) the facial expressions of Japanese and American observers watching different films. Both groups showed the same facial actions, however as predicted by different display rule of two cultures when a person in authority was present, the Japanese subjects smiled more and showed more control of facial expressions than the Americans (Ekman P., 1992) (Ekman, 1982).

Human face during face-to-face communication rarely remains still. People show a large variety of facial expressions, but not all of them correspond to an emotion and sometimes the facial expressions fulfill other communicative functions. For example raising eyebrows most of time shows the emotional state of surprise but often used as conversational signal. Other conversational signals are eye blinking or head movements. Regulators are also facial expressions which don't show any emotions but control the flow of conversation by helping in interaction between people. Eye contact and head movement are examples of regulators. Sometimes the facial expressions satisfy the biological needs of the face, like blinking to keep eyes wet and moisturizing the lips.

1.3. Problem overview

Human face conveys lot of information about our needs and feelings. People express their feelings using facial expressions. Nowadays most of our daily activities involve computers and we get more and more dependent on our computers. But the human-computer interaction is still dominated by keyboard and mouse. We should follow some sequences of instructions and learn the machine language. There is no room for human emotions and it is not a natural way of communication for human. Humans use lot of modalities in face-to-face communications and it is easier for them to communicate with computers in the same way as they communicate with other humans. Hence, as soon as computers start to become multimodal communication devices, the need for robust facial expression analysis, vocal affect analysis, speech recognition, (body) gesture recognition and context awareness became apparent. A multimodal communication system enables the computer to adapt itself to the situation and the needs of users at that moment. In the same way as it happens when we talk face-to-face and lot of information is extracted from the context and the emotions of the speaker (Pantic M., September 2003.).

Allowing input in form of speech, gestures and facial expressions and getting output on a graphical user interface and sometimes in the form of an embodied agent instead of using the traditional method (keyboard and mouse) makes it easier and more natural for human, especially when his emotions are more involved.

A multimodal communication system enables the computer to adapt itself to the situation and the needs of the users at that moment. Lot of areas in our communication methods need improvements. Here we discuss some examples of social relevance of a multimodal communication system between human and computer. A multimodal communication system enables elderly people and people with some handicaps to participate in our new communication world. A well-designed multimodal application can especially be used by these groups. Because of their handicap they rely on one of their senses more than the other and their way of interaction is far different from using the keyboard and mouse. Searching for answers and information, these days, happens online most of the time. But the current search engines don't understand the context of the command and cannot adapt the results to the wishes of the users. Using these engines can be time consuming and they don't give us the correct results all the time. Multimodal inputs/outputs can specify our questions and help us to get better and faster answers. The social networking websites like 'Face book' and 'Hyves' are very popular these days. People try to connect and interact with each other through these websites. But the communication is still not multimodal and lot of improvement is needed. These reasons and lot of others give us the indication to look for multiple modes interaction as it happens in the face-to-face communications.

Many different prototypes of expression recognition tools have been developed using visual or prosodic features. However, it remains very difficult to compare the performance of these prototypes due to the lack of common databases and protocols. Machine understanding of expressions could revolutionize human-machine interaction and has, therefore, become a hot topic in computer-vision research.

1.4. Research goals

Our main goal during this project is to extract facial emotional expressions from video recordings in face-to-face communications. To realize this we start with the following steps:

1. Start a research methodology.
2. Create a database of face-to-face communications.
3. Localize facial expressions in video recordings from our database.
4. Label the facial expressions in the recordings.
5. Track the changes in contour of eyes, mouth and eyebrows.
6. Model the human face-to-face communication.
7. Model an automatic facial feature extraction tool.

To reach our goals we perform some actions. In the table 1 we give an overview of our actions and reached goals.

Table 1: Actions and goals.

	Action	Goal
1	Research methodology.	State of art in visual and audio database, multimodal annotation system and multimodal communication systems.
2	Create database.	Collecting the face-to-face communication recordings in a suitable environment.
3	Annotate the recordings.	Producing an annotate corpus of facial expression, with defined on-set/off-set.
4	Analyze the annotated corpus.	Comparing the different presumptions of annotators and calculate the level of agreement between the annotators to find out how do we express and perceive an emotion.

5	Track points on the contour of eyes, eyebrows and mouth.	To extract the facial features for each emotional state and defining the emotion clusters.
6	Model the human face-to-face communication.	Find the triggers of human's emotional expressions.
7	Design a model.	For automatic facial feature extraction from face-to-face communications.

To reach our goals for extracting emotions from face-to-face communication, after some research methodology and gaining knowledge about the subject, we decided to build a database of different emotional states. Needless to say that data corpora are an important part of any emotion recognition study and having a good data corpus is of great help for us and other researches in this field. There are a number of data corpora available in the scientific community, however these are usually very small and tailored to a specific project. More about these databases is in chapter 3. So we decide to build our database which covers as many aspects of emotion recognition as possible. We asked native Dutch speakers to sit in front of our cameras and express emotions by facial expressions. We used two high speed cameras and two professional microphones. The cameras were controlled by the speaker, through software. This provides us a better control of recordings.

After collecting the data we start an experiment with 3 observers (non professional annotators) to annotate the recordings using our annotation protocol and an annotation tool called ELAN. We compare the different presumptions of all annotators and calculate the level of agreement between them. Calculating the level of agreement helps us to find out how much humans differ from each other in perceiving and expressing an emotion. Among others we also improved a tool for extracting the facial features from annotated recordings. This tool enables us to track the face and define the facial features for each emotional state from the sequence of images. For an easier emotion extraction we used the marked facial features from our recordings. We track points on the contour of eyes, eyebrows and mouth. Using this tool enables us detecting and recognizing the features for each emotion.

Finally we use the extracted features to build a model to extract automatic emotions from face-to-face communications. However for building such a model we need lots of correct annotated data and lots of facial features.

Special research questions in this proposal are:

- Can we successfully create a protocol for the creation of a multimodal database?
- Can we compare the different presumptions of all annotators and calculate a level of agreement between them?
- Can we track the points on the contour of the lips and eyebrows to extract facial features per each emotion?
- Can we define some emotion clusters using the extracted features?
- Can we build an automatic emotion extractor for the face-to-face communication?

Chapter 2

2. Literature survey

Need for developing an interactive communication system between human and computer is still one of the biggest challenges of our time. These communication systems enable human to communicate multimodal with computers using facial expressions, speech, voice tone, gestures and etc. For achieving a better result in an effective communication system between human and computers the automatic learning from huge data is a pre. Indexing and retrieval tool for digital media are an active research area at the present time. To understand the human's communication methods we can start with annotating the different recorded activities of people and describing in detail what happened during the communication. This way we produce an annotated corpus of the human interaction methods and create enough dataset to train the new communication systems. It helps to learn more about how people interact for developing an automated human behavior detection system.

2.1. Annotation tools

To find the best annotation tool we did a very breadth research. First we start with low level annotating tools. We got our inspirations from Benesh notation movement (Benesh).

BMN is a written system for recording the human movements and is mostly used in recording the dance works. The notation is written on a five-line stave that is read from left to right and from top of the page to the bottom. It is a 3 dimensional representation of the dance movements and helps to analysis the dance works. BMN was the source of inspiration for us to develop a notation for the movements of different elements on the face for each emotional state.

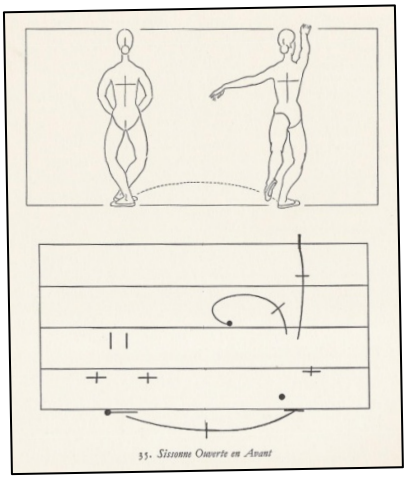
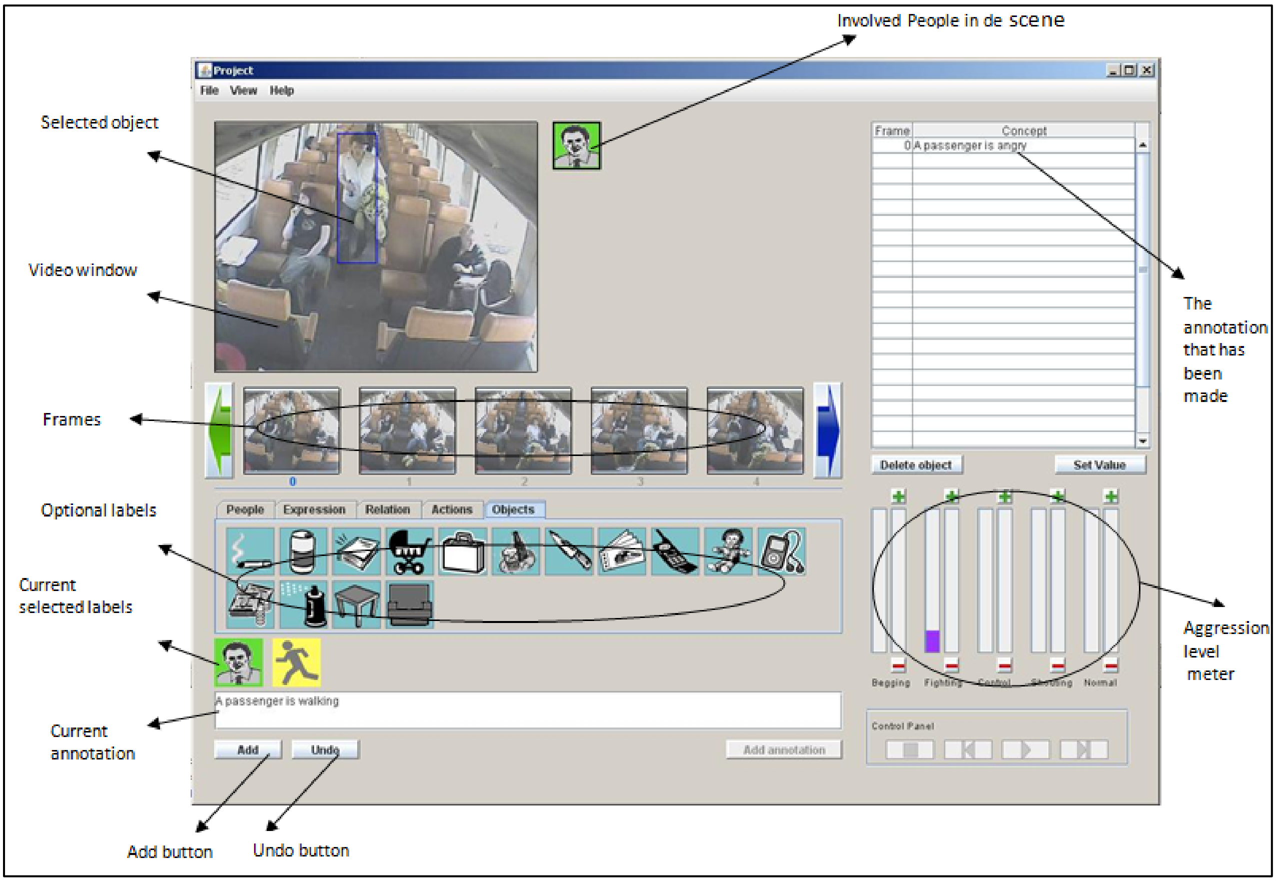


Figure 3: An example of Benesh movement notation.

As it was done by BMN we also tried to notate the movements of the different elements on the face. Notations on each line indicate the movement of each element on that line. We recorded all the movements of the important elements on the face during expressing an emotion. We used an annotation tool made by a student at TU Delft. It was an annotation tool for multimodal recordings of aggression scenarios for a train environment and returns a certain aggression level as an output. Input to this annotation tool was a sequence of images that are captured from video data or recorded scenarios of aggressive and non aggressive situation. User selects an object or person on the screen and assigns a label to it. Interface is connected to a rule based expert system to handle the incoming data from the annotating process. The output gives an aggression level and an aggression classification (Ismail, 2007). A screenshot of this tool is shown in figure 4.

But after some experiments we noticed that this is not a very user friendly way to annotate the emotions. It was a complicated way to annotate the emotions which takes lots of time and it doesn't give the exact result we were looking for. Our intention was to make an easy, clear and user friendly protocol for annotating the video recordings. But there is no need to annotate all the movements on face. They are already defined in the most commonly used system for measuring and describing facial behaviors in the Facial Action Coding System (FACS) by Ekman (Friesen, 1978). Ekman and Friesen developed the original FACS in the 1970s by determining how the contraction of each facial muscle (singly and in combination with other muscles) changes the appearance of the face. This facial activity is described in terms of visually observable facial muscle actions i.e. action units, AUs. With FACS, a human observer decomposes a shown facial expression into one or more of in total 32 AUs that produced the expression. As an addition 12 dispositions have been added to the FACS, these include the head and eye movements. FACS is a structure-based coding, closely connected to the anatomy of the face. The obtained facial expression scoring is universal across a broad spectrum of faces. Therefore FACS is widely used by psychology researchers, and it is also very common among researchers that work with facial expression analysis by machines.



Frame	Concept
0	A passenger is angry

A passenger is walking

Figure 4: Different components of an annotation tool to detect aggression in trains.

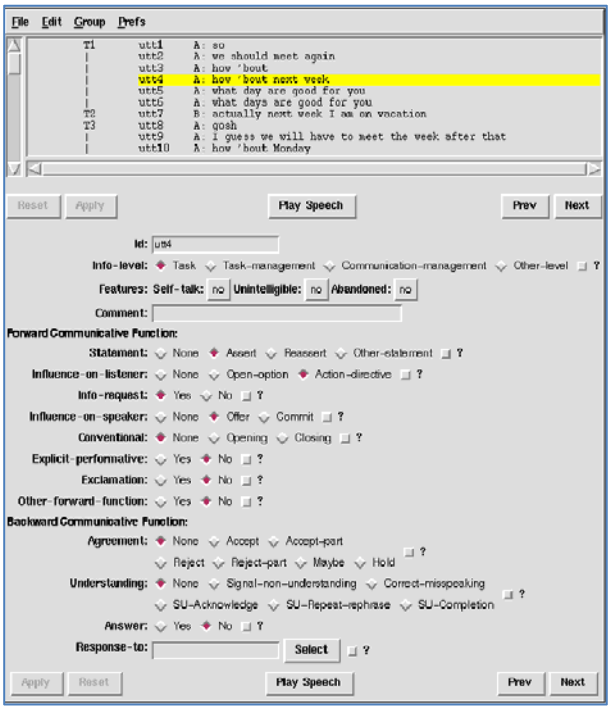
After some more researches we found some annotating tools which are used in lots of multimodal annotation projects. One of them was Anvil (Anvil). Anvil offers frame-accurate and hierarchical multi-layered annotation. It enables user to use his annotation schema using different color coding on different layers. Another multimodal annotation tool is called ELAN and is a professional tool for the creation of complex annotations on video and audio resources and enables users to annotate an unlimited numbers of annotations to audio and video streams (ELAN). It handles multiple videos up to 4 and has no practical limit on file size or duration of the videos. In chapter 4 we explain more about ELAN and the experiment we have done using this tool.

2.2.The MATE Annotation workbench

Many people wish to annotate spoken dialogue corpora with coded information. This information can come in many forms for many different purposes like representing part of speech information or syntactic structure (Klein, July 1998). It may also be non-linguistic like information about the communicative situation. Annotation task as we know is very expensive and time consuming. The MATE workbench is intended to address the need for technology by providing a single interface to all of basic functionalities which corpus annotators need, but with enough flexibility that different projects can provide different kinds of annotation and that information can be ported between the workbench and other applications.

The most important obstacle during designing an annotation tool is the fact that corpora are not necessarily hierarchically arranged. It makes it difficult to design an efficient algorithm which also acts flexible enough to easily implement a new annotation scheme. Each type of structure can refer to some shared base level of transcription like orthographic or phonetic or it might involve several kinds of tags like dialogue moves, phrases and sentences. To start with designing a general annotation tool it is much easier to start with implementing a generic solution which can be adapted to other tag sets. The existing tool of MATE called DAT tends to support particular coding scheme and allows tags to be renamed or extra categories to be added to a tag set. Other existing tools have just fixed methods for displaying information and fixed actions for coding. They are often built for particular task and have a fixed user interface. MATE Workbench aims to improve the existing tools by representing more flexible solutions for annotating corpus.

DAT was developed at the Department of Computer Science University of Rochester, Rochester, NY, USA, by Mark Core and George Ferguson. It is a coding tool and display facilities for a range of coding schemes at various levels of annotation from prosody to dialogue acts and communication problems. These annotation schemes are being chosen based on a wide review (Klein, July 1998). The input formats of the dialogues uses markup tags to store information and each file contains a whole dialogue. Each file should be segmented before annotation.



```
T1    utt1    A: so
|     utt2    A: we should meet again
|     utt3    A: how 'bout
|     utt4    A: how 'bout next week
|     utt5    A: what day are good for you
|     utt6    A: what days are good for you
T2    utt7    B: actually next week I am on vacation
T3    utt8    A: gosh
|     utt9    A: I guess we will have to meet the week after that
|     utt10   A: how 'bout Monday
```

Id: utt4

Info-level: ◆ Task ◇ Task-management ◇ Communication-management ◇ Other-level □ ?

Features: Self-talk: no **Unintelligible:** no **Abandoned:** no

Comment:

Forward Communicative Function:

Statement: ◇ None ◆ Assert ◇ Reassert ◇ Other-statement □ ?

Influence-on-listener: ◇ None ◇ Open-option ◆ Action-directive □ ?

Info-request: ◆ Yes ◇ No □ ?

Influence-on-speaker: ◇ None ◆ Offer ◇ Commit □ ?

Conventional: ◆ None ◇ Opening ◇ Closing □ ?

Explicit-performative: ◇ Yes ◆ No □ ?

Exclamation: ◇ Yes ◆ No □ ?

Other-forward-function: ◇ Yes ◆ No □ ?

Backward Communicative Function:

Agreement: ◆ None ◇ Accept ◇ Accept-part ◇ Reject ◇ Reject-part ◇ Maybe ◇ Hold □ ?

Understanding: ◆ None ◇ Signal-non-understanding ◇ Correct-misspeaking ◇ SU-Acknowledge ◇ SU-Repeat-rephrase ◇ SU-Completion □ ?

Answer: ◇ Yes ◆ No □ ?

Response-to: □ ?

Figure 5: Annotation tool DAT.

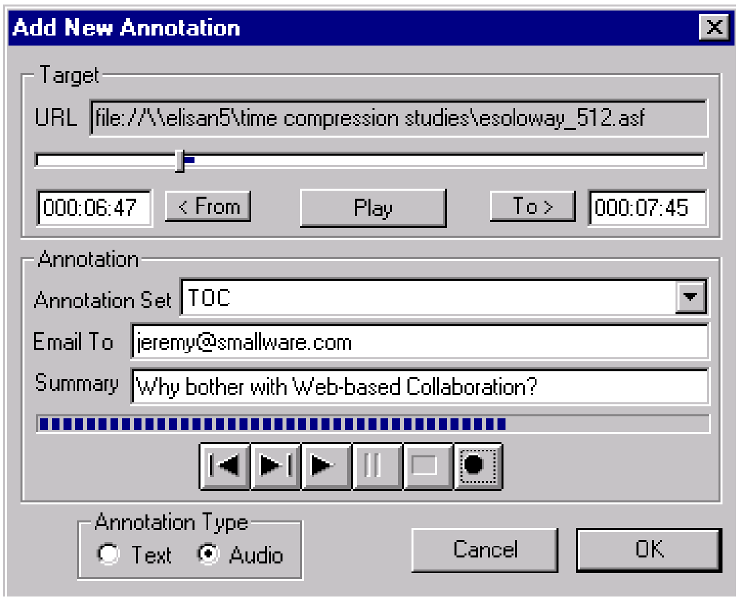
When user selects a segmented line a new window will appear and enables user to enter the information. On this window the annotation dimensions are presented in boldface on the left side of the window, and the possible values on the rest of line and the selection is made by pressing the small button to the left of the label.

The annotation scheme of DAT has 15 different dimensions and user has to select one tag for each annotation dimension, requiring at least 15 clicks per segment to annotate. The large amounts of clicks and bad designed buttons made the interface weak and not a good example for a productivity solution in the area of annotation.

2.3. Microsoft Research Annotation System (MRAS)

The use of streaming video on the Web for workplace training and distance learning has generated a lot of discussion and ideas lately. The ability to view content on demand anytime and anywhere could expand education from a primarily activity to include more flexible, asynchronous interaction. For example it allows students to record questions and comments as they watch Web-based lecture videos. It can allow them to share annotations with each other and instructors, either via the system or via email. It can enable them to watch their own and others' annotations scroll by as the video plays. It can support using annotations as a table-of-contents to allow jumping to relevant portions of the video. And finally, such a system can allow students to organize their annotations for personal and public use. (Bargeron D., 1999)

MRAS is a prototype video annotation system built for adding text or audio annotation to the video streams on a web page. When a user wants to add an annotation to a video recording , using the 'add' button on the interface opens a dialog box. It supports adding text and audio annotation. The annotation's target position is set to the current position in the video's time line.



URL file://\\elisan5\time compression studies\esoloway_512.asf

000:06:47

000:07:45

Annotation Set TOC

Email To jeremy@smallware.com

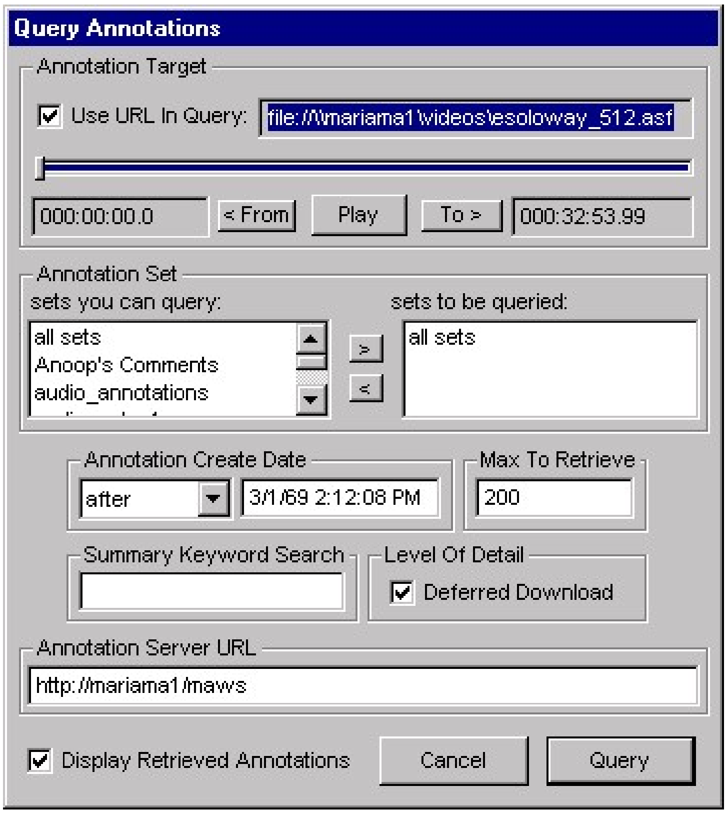
Summary Why bother with Web-based Collaboration?

Annotation Type ○ Text ● Audio

Figure 6: Adding an audio annotation to a video recording using MRAS.

In figure 6 is an image of interface of prototype MRAS shown. In this example an annotation is recorded and added to an annotation set called 'TOC'. The annotation is also mailed to Jeremy@smallware.com. The annotation is contextualized by the target video from 6 minutes and 47 seconds to 7 minutes and 45 seconds.

As adding annotation to a recording it is also possible to retrieve an annotation from the MRAS server. The dialog box is accessed via the ‘Query’ button and it enables users to search for annotations. Later if the user wishes to see the full content of a particular annotation he/she can download the data he/she is interested in, without downloading the content of other annotations. In figure 7 is ‘Query Annotations’ box shown. Using this box user can decide which part of annotation he/she is interested in. The decision can be made based on the time of the annotation or the annotator who made the annotations.


Query Annotations

Annotation Target

☑ Use URL In Query: file://\\mariama1\videos\esoloway_512.asf

000:00:00.0 — 000:32:53.99

Annotation Set

sets you can query: all sets, Anoop's Comments, audio_annotations

sets to be queried: all sets

Annotation Create Date: after 3/1/69 2:12:08 PM

Max To Retrieve: 200

Summary Keyword Search:

Level Of Detail: ☑ Deferred Download

Annotation Server URL: http://mariama1/maws

☑ Display Retrieved Annotations

Figure 7: User can use the 'Query Annotations' box to search for a annotation.

The fact that web-based annotation can be shared has a big impact on users. The user was asked if the sharing helped them or distracted them. The majority of the users found others’ annotation useful in guiding their own thinking, but some of them reported the additional information confusing. They were also asked whether they found using MRAS to be an effective way of preparing for a class discussion (Bargeron D., 1999).

A strong majority across all conditions agreed. The average response of subjects in the Text and Audio condition was 6.0 out of 7 (with 7 "strongly agree" and 1 "strongly disagree"). The average in both the Audio-Only and Text-Only conditions was 5.0 out of 7.

2.4.SmartKom

SmartKom is a multimodal dialogue system that combines speech, gesture and facial expression for input and output. SmartKom is the follow-up project of Vermobil and uses some of components of Vermobil. The goal of Vermobil project was developing a (prototype of) translation machine, combining two technology, speech processing and machine translation for finding a date for a business meeting.

SmartKom provides an intelligent computer-user interface which allows the user communicate with computer on the human natural style of communication. The system is also able to recognize the gestures and the speech of the user. User delegates a task on an anthropomorphic which is a graphical representation of the user as an interaction agent on the display. The interaction agent analyzes the facial expression and recognizes the speech of the user. He has access to different IT services and search for results behalf of the user and finally represents the output which matches the emotion and the wishes of the user. This agent called Smartakus and he has an “i” shape which reminds of the “i” that is often used as a sign for information kiosks.

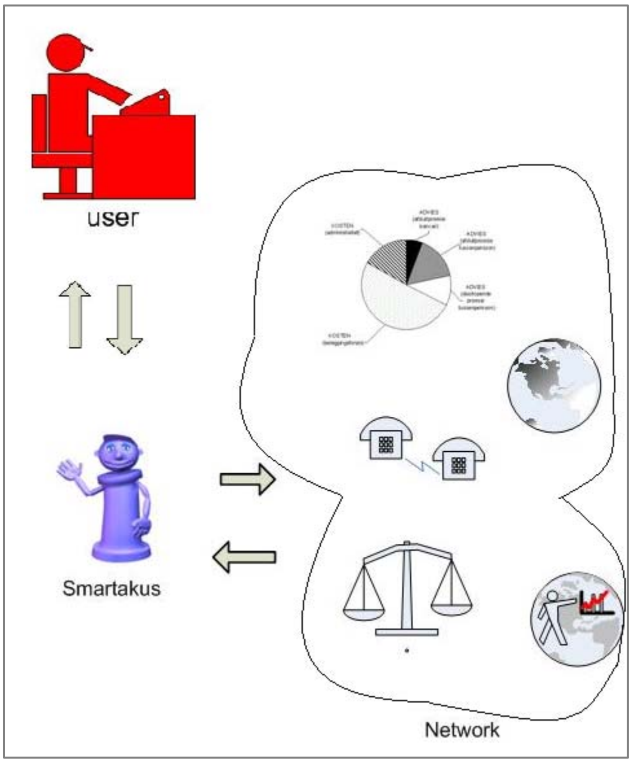


Figure 8: Delegation-oriented dialogue paradigm of Smartakus.

For a truly multimodal communication SmartKom has three different user interfaces. Spoken dialogue, graphical user interface and gestural interaction and enables the user to communicate through his natural method of communication. The graphical user interface instead of using the traditional WIMP (*windows, icons, menus, and pointer)* uses the natural gestural interaction of the user combined with facial expression. For this purpose is an extended version of SIVIT (Siemens Virtual Touch Screen) used, which recognizes real time facial expressions. For representation of the information that transfers between the components of the system they use a XML-based markup language, called Multimodal Markup Language or M3L.

Smartakus has 3 different user interfaces:

- SmartKom public: It is a multimodal communication kiosk for public places like airports, train stations, hotels, restaurants, theaters and etc. Facial expression of the user is captured using a DV camera, the gestures are tracked with an infrared camera and the speech is captured with a directional microphone. The system projects the output on horizontal display and speech of Smartakus is provided through two speakers under the projection surface.

- SmartKom home/office: This interface of Smartakus uses a virtual touch screen and his associated software is able to recognize real-time facial expressions and speech. Smartakus home/office provides electronic program guides for TV, controls consumer electronics devices like VCRs and DVD players. It also has access to standard applications like phone and e-mails.

- SmartKom mobile: This version of Smartakus uses a PDA front end. The mobility is one of the key concepts of SmartKom mobile and it is a constant mobile companion who can be used in cars as the navigation system and route planner or just be carried by a pedestrian walking around on the street and provides personalized mobile services on street. It gets gestures and audio inputs from user and provides output as audio and information on display. Access to internet is provided via a GSM connection.

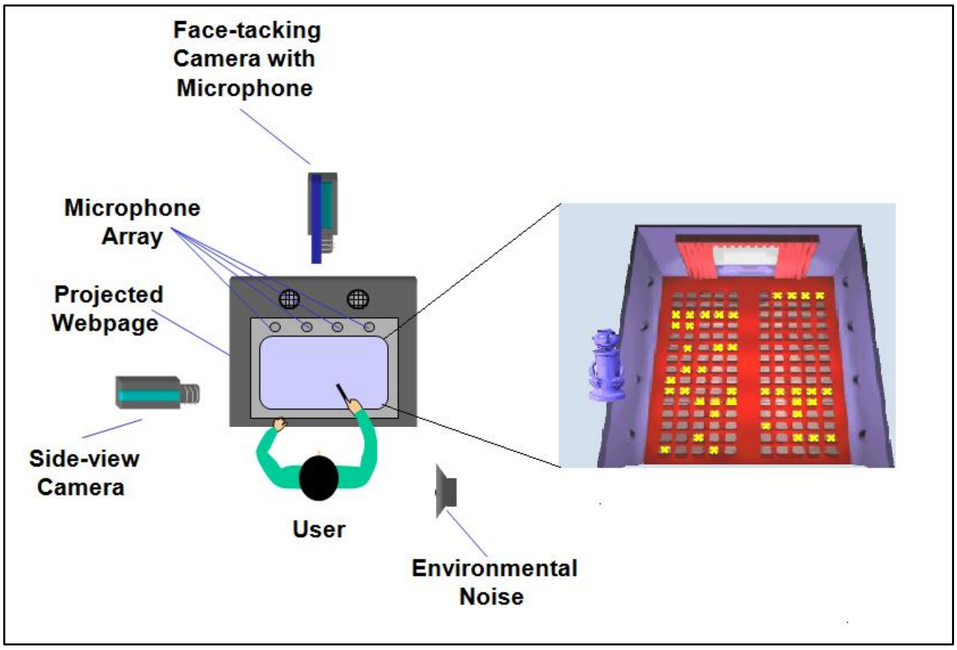


Figure 9: Multimodal input and output in Smartkom. Smartakus helps a user with the cinema ticket reservation.

Smartakus accepts input in different modalities, in form of natural speech, gestures and facial expressions. The prosody of speech and facial expression is analyzed carefully. During the project there are lots of data collected from different test persons. The collection of data is done with Wizard of Oz technique. The test person interacts with a system which is simulated by two humans from another room. The speech, facial expression and gestures of the test persons are recorded with two microphones and two digital cameras. Also the hand gestures are captured with an infrared sensitive camera. Each recording was about 4.5 minutes (Steininger S., 2006). After collecting the data, it is labeled for training the recognizer as well as for developing a model to predict typical human-machine interactions.

The user-states are defined with regards to the subjective impression that a human partner would have and it doesn't use a specific system like the FACS. In FACS the precise morphological shape of facial expression is coded but in SmartKom is a simplified, practice-oriented system used, which doesn't code the user-state but the impression that the communicated emotion or state generates.

Because it is still unknown which feature of the face and voice contribute to an emotional impression and which degree of each feature contribute to the impression, defining the coding convention that marks these feature in the data was not possible. That is why they used another strategy for labeling. In their strategy the labelers mark the beginning and the end of a user-state sequence and sort it into one of several subjective categories.

Labeling is first done with some defined subjective categories. These categories are 'anger', 'boredom', 'joy', 'surprise' and 'neutral'. The beginning and end of the user state were defined by an observable change in emotional state. The weakness of the user-state was also marked. But later on there were some changes made to these categories. For example the category boredom was excluded, because it was not really recognizable from neutral state. Also category 'anything else' was added for the states that did not show a neutral expression but also no meaningful label could be given to them. There are another two new categories added, 'helplessness' and 'pondering'. Next step was finding the detectable feature. For each category there are some characteristic feature listed. The feature should occur regularly or it should be very distinctive of a category.

For semantic representation of all the information flowing between the components of SmartKom is M3L used. M3L (Multimodal Markup Language) is an XML-based language used for exchanging and transferring data between al components.

Smartakus is not available yet for public use. There are just only some demos on the SmartKom website (Smartkom) which demonstrates different modalities of Smartakus. User talks to Smartakus and uses gestures. SmartKom home enables user to control al the electronic devices at home, like TV, VCR and DVD player and also gets the weather service and traffic information. When Smartakus discovers a new output device he switches to the more suitable device and it does not interrupt the task that he was doing. Not all the SmartKom's applications use the same modalities and sometimes it is better to switch between the applications. This way user can continue with the same task that he already was doing without being interrupted on the other application. When user is walking out of the house and getting into his car, Smartakus is able to switch from SmartKom

Home to the SmartKom Mobile and continue with the task which he was doing. For some other tasks which more multimedia devices are needed using SmartKom home is more skilful.

As mentioned before SmartKom uses gesture and speech as input. User's face is detected with a camera and based on the detected user's facial expressions SmartKom recognizes if the user is satisfied or not. The hands' gestures are also captured by an infra red camera. For getting access to the personal account user is identified through a biometric identification process. It involves one of the following identification processes, identification of the voice, signature or the hand. Once the user is identified he gets access to the system and to his own data which makes it very personal for the user.

2.5. Ruth

Ruth, the Rutgers University Talking Head is a real-time animated talking head which is able to reproduce the functions and behaviors of a natural face-to-face conversation. In developing Ruth is tried to make a conversational agent which is able to express emotions through facial expressions and speech. Ruth has a well designed head. His facial expressions are made as a combination of motion's units and prosodic motions. It is a typical parallel between verbal and nonverbal channels and produces facial displays and head movements in synchrony with speech and lip movements (DeCarlo D., 2003) (DeCarlo D., 2004). Ruth accepts his input as the symbolic description of conversational behavior. In Figure 10 a screenshot of Ruth is given. Using the menu on the right side of the page you can choose a sentence and Ruth would read it for you and shows the matching facial expressions.

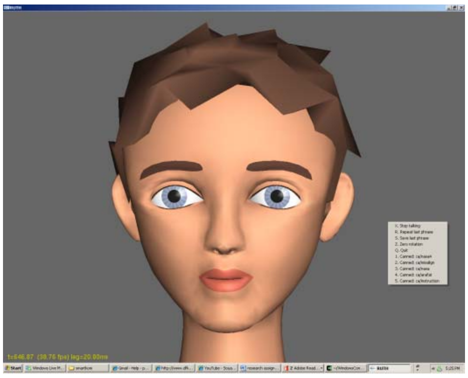


Figure 10: Ruth, the talking head.

Facial expressions of Ruth are expressive enough to show his emotions clearly. There are two different video's made, with exactly the same soundtrack and two different visuals (http://www.cs.rutgers.edu/~village/ruth/research.html). The scenario of these two videos is about a host and his guest. The host served just chocolate desert as dinner. In the first video, Ruth (the guest) shows the polite reaction for choice of desert and in second one he is a little bit sarcastic and not so satisfied. These two different emotions are completely visible through his facial expressions. In both videos there is just one soundtrack used but two different visuals. Watching them, based on the expressions that Ruth shows we understand directly that he expresses different emotions in each video (Stone M., 2003).

Ruth is a freely-available cross-platform real-time facial animation system that animates high level signals in synchrony with speech and lip movement (DeCarlo D., 2004). RUTH uses markup text with synchronized annotations for intonation and facial movements, including eyebrows, eye blinks and head movements (Marsi E., 2007). For producing speech it uses festival text-to speech system (Black, 2002). He relies on ToBi (Tone and Break Indices) systems, a framework system for transcribing a language intonation. ToBi is not an international phonetic alphabet for prosody, each ToBi is specific to a language variety and the community on that language variety (ToBi). The intonations and prosodic organizations differ from language to language, and often from dialect to dialect within a language. In the Dutch version of Ruth, the makers of the Dutch Ruth use the equivalent system for annotating Dutch intonation instead of original intonation system (Gussenhoven C., 2005). It is known as ToDi (Transcription of Dutch Intonation) and is supported by Nextens TTS system for Dutch (Marsi E., 2007).

In developing Ruth is tried to make a conversational agent which is able to express emotions through the facial expression and through the speech. He accepts his input as the symbolic description of conversational behavior. Hereunder a sentence of Ruth's input as an example is given. It presents the annotation of a group of four analysts. The intonation is specified according the *Tones and Break Indices* (ToBi) standard (Silverman K., 1992) (Beckman, 1997).

TEXT	*far*	*greater*	*than*	*any*	*similar*	*object*	*ever*	*discovered*
INTONATION	**L+H***	**!H*H-**			**L+H***	**!H*L-**	**L+H***	**L+!H*L-L%**
BROWS			**[**		**1+2**		**]**	
HEAD	**[**	**TL**	**]**		**D***		**U***	

The first line contains the text. On the second line we have intonation. L+H, !H* and L+!H* mark accents on syllables while H-, L- and L-L% record tones at the boundaries of prosodic units. The third line, 1+2 indicates the movements of brows, according to the Ekman coding system (Ekman P., 1979). The head movement is also labeled and shown in fourth line, TL for the tilting nod on a phrase, D* for downward nod and U* for an upward.

Ruth uses tagged text to produce the animations. In general such tagged text is obtained using heuristics to produce elaboration on plain text input as is done in BEAT system (Behavior Expression Animation Toolkit). BEAT allows animators to use typed text that they wish to be said by an animated human figure as input. It gives an output as an appropriate and synchronized nonverbal behaviors and speech. These outputs are sent to different animation systems (Cassell J., 2000). Ruth uses Festival speech synthesis system which works almost the same as BEAT. It drives sounds and timing file for animation in the Festival formats and enables RUTH to take tagged text and realize it as animation. In this format an utterance begins with an open parenthesis and ends with a close parenthesis and has a space-delimited sequence of word-elements in between. A word element can be a single word without any accent or boundary tone or it could be a pair, a list with two elements. The first element of the list is the word itself and the second element is a list of attribute values. It could be any specification to modify the pitch accents, boundary tones and facial actions. Using the utterances he drives sounds files and animation instructions for corresponding lip, face and head movement.

The head movements and the eyebrows movements of Ruth make hem looks more natural. The movements of his lips also match the words he says. Making Ruth could be a good begin for building interactive conversational agents for face-animation research. Displaying the appropriate non-verbal behavior enhances the interaction between men and the embodied agent.

Chapter 3

3. Data collection

For building a multimedia communication system to extract emotions from face-to-face communication in video recordings, automatic learning from huge data is a pre. But first we should manage to obtain a good data corpus (well designed, capturing both general and also particular aspect of a certain process). It is of course always possible to create our own database. But before start collecting the data several issues should be solved like do we use an existing database or we build our own database? Do we use posed data or spontaneous data? What are the environment conditions?

There are two methods known for collecting the data. Recording spontaneous data or acted data. Using human-human dialogs from TV interviews provide us spontaneous data. Some of accompanying problems of the TV interviews are, lighting problems or orientation problems. As the emotions are expressed in the real situation it is difficult to get the subjects and their faces always right in front of camera and subject may turn his/her face in all the direction (Ekman P., 1992). Wearing glasses or having beard or moustache are some other obstacles for tracking the movements of eyebrows and mouth as it is covered under glasses and facial hairs. Recording people's activities in their daily life with a webcam or a camera also provides us spontaneous emotions, but human shows limited number of emotions in daily life and putting the subjects in the emotional situation to arouse them or make them afraid or surprise is not ethical responsible.

Ekman uses this method during his study to acquire data. He showed some stress including film and picture of surgeries to get disgusted look on the face of participants for collecting data for his database (Ekman P., 1992). These methods may help us to get the real facial expressions but they are not the practical way to collect data and have some ethical problems because of putting the subjects in such emotional states. Because of mentioned problems some of the researchers prefer using acted data. In the acted scene the subject is asked to sit in front of camera and express the asked emotions in the camera direction as it is done in Cohn Kanade database. This image database consists of approximately 500 image sequences from 100 subjects. Accompanying meta-data include annotation of FACS action units and emotion-specified expressions. Subjects range in age from 18 to 30 years. Sixty-five percent were female, 15 percent were African-American and three percent Asian or Latino. The observation room was equipped with a chair for the subject and two cameras, each connected to a video recorder with a synchronized time-code generator. Subjects were instructed by an experimenter to perform a series of 23 facial displays that included single action units (e.g., AU 12, or lip corners pulled obliquely) and action unit combinations.

This database and other collected acting databases give us less real emotion and the acted expressions may differ in appearance and timing from spontaneous occurring emotions. Normal people and non professional actors are tens in front of camera and don’t show their real emotions. Some researchers also tried to record the subjects while they were playing some game with computer but as the subjects were interacting with computer they didn’t show many emotions.

After studying acted and spontaneous data, researchers concluded that expressions can be convincingly portrayed by professional actors and some non professionals (Ekman P., 1992). They found some similarities in acted data which suggest that the use of acted data is allowed. Despite all the mentioned obstacles to collect satisfying acted data it gets more preferences than using the real data.

For our project we decided to build our own database using professional actors. It is filled with video recordings of different facial expressions which also can be used by different research groups. In the following part we explain how do we collect the data and how does our database look like.

3.1. The New Delft University of Technology Audio-Visual Speech Corpus

During this project we tried among others to build a database (The New Delft University of Technology Audio-Visual Speech Corpus) which can be used by different research groups. We start our recordings by asking some Dutch native speakers to sit in front of camera and to get them in the good mood for expressing emotion we asked them to read a short story. Each participant should express 21 different emotional states (Desmet P. M. A., 2002) and each emotional state was provided with a short story called scenario. Their reaction was recorded by two sensitive microphones and two high speed cameras.

We made the recording in our lab at 12th floor of EWI building (Chitu A., 2008). In the lab the participant take place on a chair and in front of the chair there is monitor about 1 meter from the participant. There are two sensitive synchronized cameras, about two meters from the participant one in front and one at the left side of the participant. The first microphone is placed about half of the meter of the participant and the second one is about two meter from hem to subtract the environment noises from participant voice.

The experiment is done in a closed room with good lighting conditions. A good lighting condition means that there is enough diffuse light to leave no shadows on the participants face. The camera is focused on the participant. The height of the camera must change for every participant in order to get a perfect frontal view, but this is easily solved by placing the participant higher or lower in the chair. The background behind the participant is covered with a dark color, preferably blue or green.

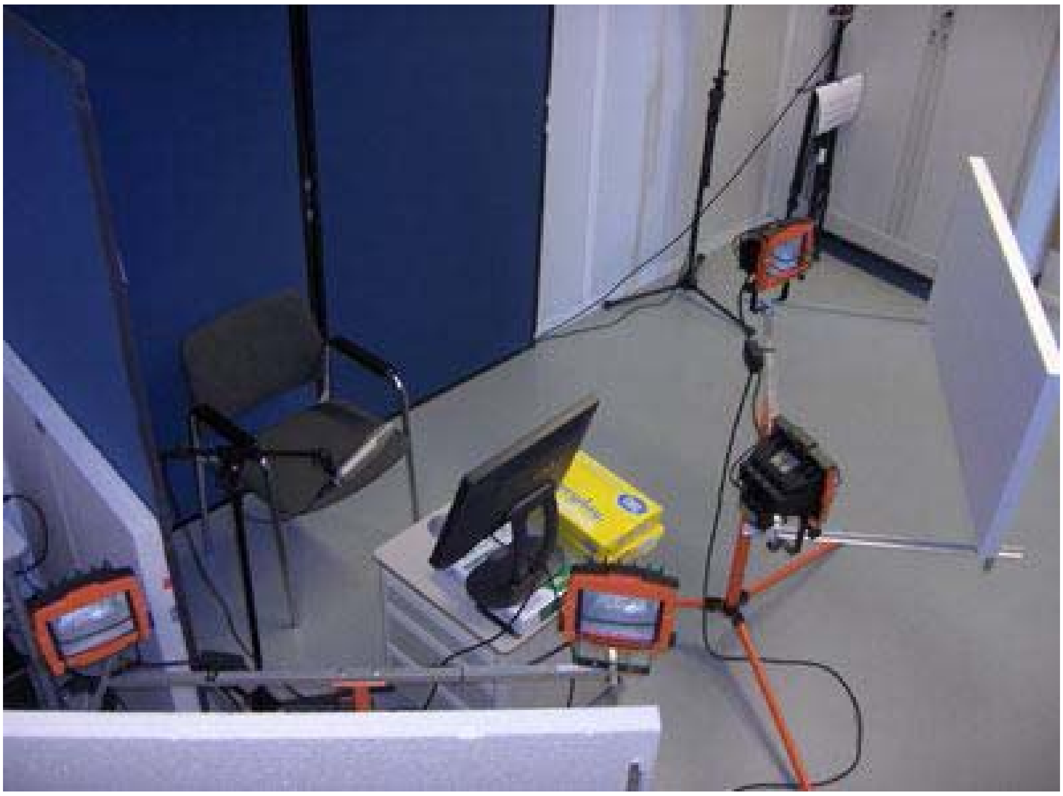


Figure 11: The setup of the recording room.

We made 8 different sessions per person and each session contains all the 21 scenarios which each scenario contains five pre-defined reaction's sentences. The reaction's sentences have a random order of sequences during all the sessions. For recordings we put 31 colored markers on the face. The colored markers are for an easier recognition but to complete our database for future research we made also all the recordings without the green stickers. There are also differences between recordings containing both facial expression and speech or just facial expressions. In table 2 is an example of a recording session is shown. The complete scenarios are in Appendix C.

The average time the experiment lasted was 45 minutes. For each emotion the participant was asked to listen carefully to a short story and to 'immerge' themselves into the situation. Once ready, the participant may read, memorize and pronounce (one at the time) the five proposed utterances, which results in five different reactions to the given situation. The participants are asked to put in as much expressiveness as possible, producing a message that contains only the emotion to be elicited. With this procedure we let our participants take multiple sessions. This way we can fill the database quickly with many different recordings. The goal is to create a balanced database with respect to gender and age. The complete recording protocol can be found in Appendix C.

Table 2: Situation and reactions to elicit 'Amusement'.

Amusement “Jij loopt met een vriendin op straat langs een kiosk. Opeens zie jij de foto van je favoriete zanger op de cover van een tijdschrift. Onder de foto staat dat hij binnenkort in een film van een beroemde regisseur de hoofdrol zal spelen. Jij vindt het werk van die regisseur ook heel goed. O, wat wil je die film graag zien.”	
Reaction R1: Dat zal een goeie film worden. R2: Die film wil ik zeker zien. R3: Hij maakt altijd goede films. R4: Wanneer gaat deze film in première? R5: Dat wordt kijken geblazen!	dAt zal en Guj@ film wOrd@ di film wIl Ik zek@r zin hE+ makt AltE+t Gud@ films wAner Gat dez@ fIlm In pr@miE:r@ dAt wOrt kE+k@ G@blaz@

After recording the scenarios, we store the recorded data for further use. Using a high speed camera increases the storage needs for the recordings. It is almost possible to record everything and then during the annotation process, cut the clips to the required lengths (Steininger S., 2001). Therefore we let the participant control the beginning and end of the recording of a clip through the mouse buttons of a wireless mouse that was taped on the arm of the chair. The result was synchronized audio and video clips already cropped to the exact length of the utterance. After a series of trials we conclude that this level of control is sufficient and not very disruptive for the speaker. The size of recording increases fast, the transfer rate of the data to the hard disk should therefore not be a limiting factor. Two high speed cameras connected to one computer gives a very large data stream, to solve the problem of writing this data to the hard disk in real-time we used two 250 GB SATAII hard disk in RAID 0. RAID stands for “Redundant Array of Inexpensive Disks”.

3.2.Anna’s recordings

As the recordings we made in our lab were short recordings of a single expression at a time, annotating them didn’t give us enough variations. To get better results and more variations we decided to use the recordings made by Anna Wojdel. Recordings made by her had almost the same settings as our recordings and she also used colored markers on face for easier tracking. These recordings were made during her PhD research. The goal of her experiment was to obtain and analyzing the recordings of people, showing facial expressions appropriate to a given situation in a face-to-face communication. Recorded persons were not professional actors. As normal people (not professional actors) do not behave naturally in front of camera and they are usually tens and show much less facial expressions than in real life she asked the participants to behave emotionally in the same way as adolescents behave while playing with children. She believes that exaggerated facial expression do not really disturb the results (Wojdel A., 2005). For recordings they used two synchronized digital cameras, one for recording the subject and one for recording the supervisor. In order to arrange the environment of recording similar to the conversation in real life, subject and supervisor were sitting (almost) in front of each other at a distance about 1.5 meter. Subject was asked to limit his/her head movements during the recordings (Wojdel A., 2005). The video sequences are sampled at 25 frames per seconds and the video resolution is 720x576 pixels.

There are 10 different scenarios which are done by 10 different persons (5 males and 5 females). Each recording is about 1,5 minute and in total about 14 minutes. Each set of recordings consists of three recordings. The first set is a subject just listening to the text and expresses the proper emotions to the text. The subject didn't see the text and there were also no suggestions given to the subject about the interpretation of the text. We refer to it as recording of listener. In this session the subject is showing an emotional state based on the scenario read by the supervisor. In the second set the subject was reading a part of a dialog and performing the role of one of the two characters. The other role is performed by the supervisor. We refer to it as the recordings of a dialog. Finally in the last set, the subject was asked to read the whole scenario and expresses the proper emotion. We refer to it as recordings of a narrator. For further analysis we select the listener recording of Anna herself. The reason to select these recordings was, as talking produces lots of unnecessary noises around the mouth, selecting the listener recordings prevent dealing with these unnecessary noises. Among all the 10 participants acting the scenarios, Anna was the most expressive one and showed much more emotional states than the other 9 participants.

Table 3: Different type of recordings and the reactions.

	Type recording	**Type reaction**
1	Listener	Facial expressions of the listener.
2	Dialogue	Facial expression and speech of the dialogue player
3	Narrator	Facial expression and speech of narrator

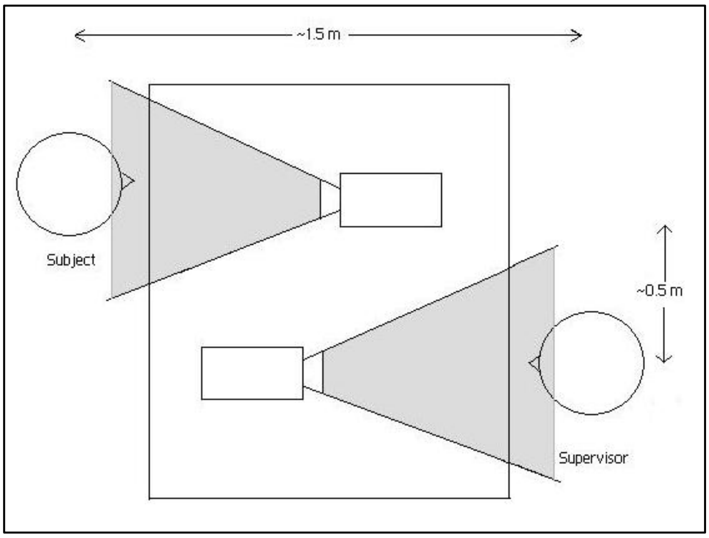


Figure 12: Set up of Anna's recording.

Dialogues used in the recordings are from popular polish juvenile book “kwiat kalafiora” written by M. Musierowicz. As this novel is intended for young people, characters appearing in the book have very distinct personalities and different temperaments. They are very expressive and their emotions are often described in an exaggerated way. In the tables 4 a section of a used scenario and translation of it in English is given. Anna selected 10 fragments, which contain dialog between two characters from the book. In each session two persons took part, a subject and a supervisor. The recordings of the subject were used for analysis and the recordings of the supervisor were used for further reference. As our goal was to analyze the facial expressions of a face-to-face communication during the recordings of the listener, listener sympathized with the both personages in the story and showed the corresponding emotions and he/she is not emotional natural. We will use these recordings to follow the movements of the mouth, eyes and eyebrows during each emotional express and to facilitate the automated extraction some makers (31 points) are put on face. Figure 12 presents, the usual appearance of “astonishment”.

Table 4: A fragment from Anna's recordings.

— Ale posłuchaj. Najpierw gdzieś z piwnicy słychać było dziwny dźwięk – jakby stukanie, czy turkot. Potem zaskrzypiały drzwi, coś głucho trzasnęło, a na koniec rozległ się blaszany, przykry łoskot.	— Listen to me. First I heard from the basement a strange noise — something like knocking or rattling. Then a door creaked, something crashed dully, and finally I could hear metallic and annoying crack.
— Żartujesz! — powiedziała tylko Ida, kiedy Gabriela skończyła zarówno opowiadanie, jak i naleśnik.	— You are kidding! — Said Ida, when Gabrysia finished both, her story and pancakes.
— Wszystko prawda. — Więc ta wariatka nie jest wariatką? — Och, Ida, Ida. Jak ty się wyrażasz,	— All of it is true. — So, she isn't crazy ? — Oh, Ida, Ida. Watch your manners,

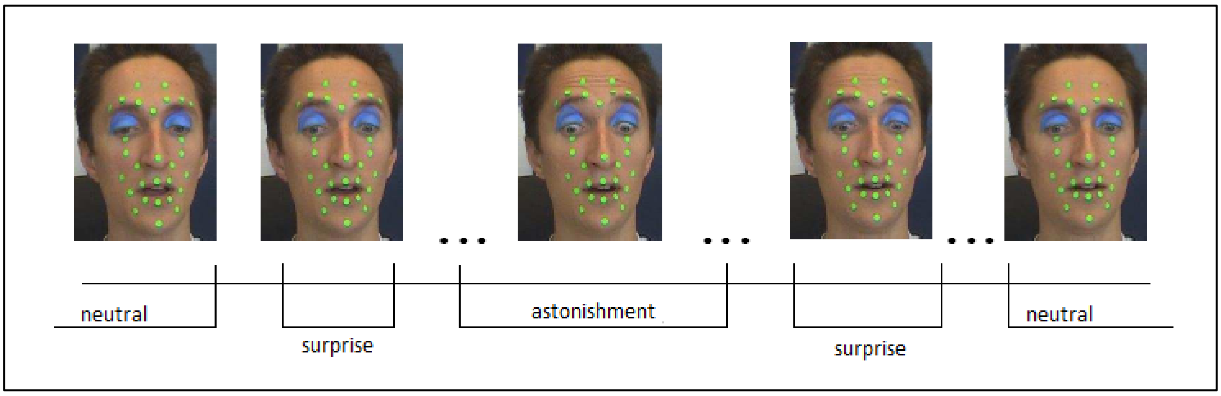


Figure 13: Typical appearance of facial expression: ‘astonishment’.

The listener shows only facial expressions. The question is what is happening if the listener starts speaking and plays the set of actor. It can be assumed that the emotional state of listener (in the role of actor but not speaking) and actor is the same. Emotional can be displayed by facial expressions and speaking style. A challenging question is if the displayed emotion by facial expression and the speaking style are the same, do they complete each other or are they different or even opposite. Next there is an interaction between facial expressions and speech. It is impossible to smile and to speak. So we can expect that the facial expressions of a listener and a speaker are different. May be the expression from the upper face is the same, but the lower area is quite different.

Chapter 4

4. Annotation

People do numerous activities during a day and almost all the activities involve communication with each other. In order to have a better understanding of human-human communication and to improve the human-computer interaction it is essential to identify and describe the different modalities of human-human communication including collection and annotation of multimodal data. Describing the different activities of people in detail is called annotation and produces an annotated corpus of recorded human interactions. Annotated corpus helps to learn more about how people interact and is also useful for developing the automated behavior detection systems. In our case we are interested in the onset and offset of facial expressions, the label of the expressions and what triggered the facial expressions. To achieve these goals we start an annotating experiment. We started our experiment with 3 different annotators. In this chapter we explain how we annotate our data and what the results are.

4.1. Experiment

To start our experiment first we had to find the proper multimodal annotation tool and an annotation method. More about the annotation tools has been explained in chapter 2. After some research we chose an annotation tool called ELAN to work with. ELAN is a professional tool for the creation of complex annotations on video and audio resources and enables users to annotate an unlimited number of annotations to audio and video streams. Depending on the task, annotation could be a sentence, word, comment, translation or description of an observing in media (ELAN). Annotations in ELAN are created in multiple layers called tiers and it is possible to navigate through the media with different step size. In our case, annotation was a chosen label of an emotion from our offered list. The list contains 12 labels made by Anna Wojdel. We made an individual tier (layer) per annotator in the program for the given labels and an individual tier for the triggers.

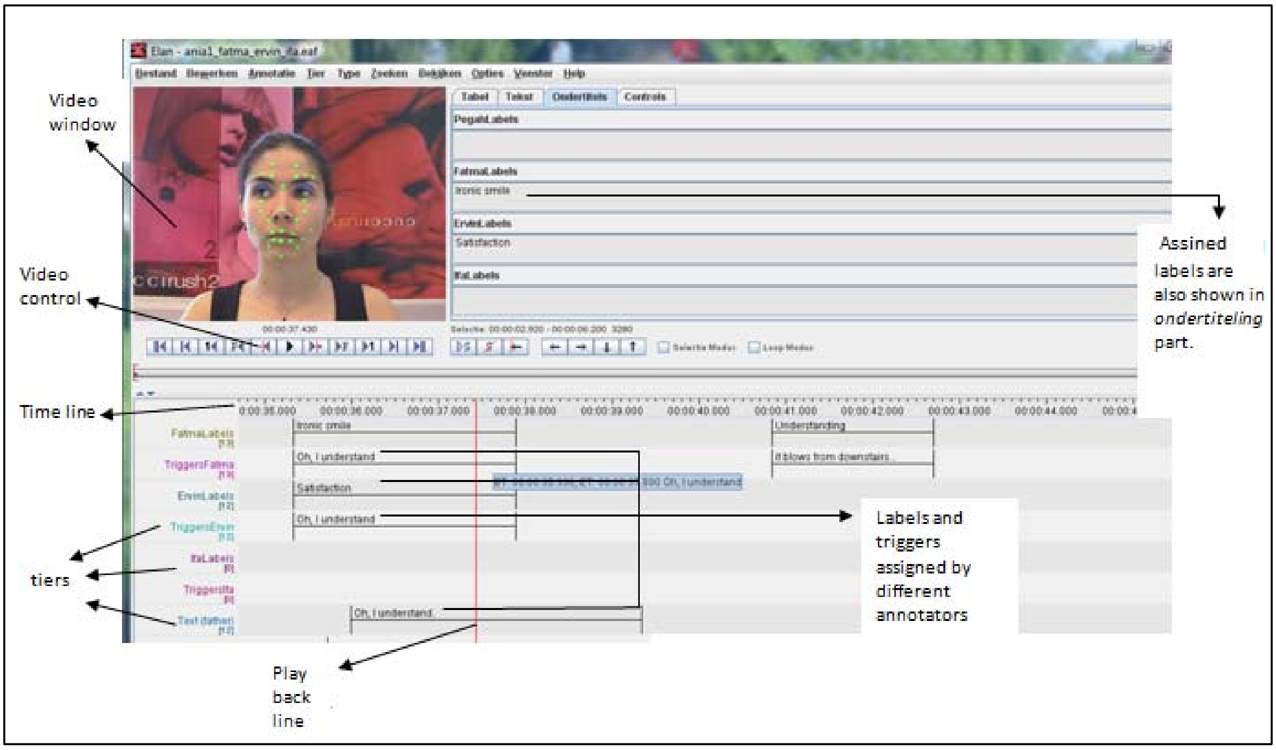


Figure 14: Annotating a recording using ELAN.

The recordings made by Anna were in Polish and as the text has a big contribution to human's interpretations, we put the translation of the text on a tier in the program. This makes it easier for annotators to understand the expressions they see and define a trigger for the label. We explain more about triggers in section 4.3. Annotators annotate the recordings according to an annotation protocol. This annotation protocol is given in appendix A. During the experiment annotators were able to discuss the observed expressions and label the expression. They do not necessarily need to agree with each other about the given labels and they were allowed to give their independent labels. If an observed expression doesn't exist on the list they were also allowed to add it as a new label to the list. During the experiment none of the annotators added any new label to the list.

To annotate the recordings, they watch the recordings and when they observe an emotional expression they assign a label to the number of the selected frames on the activated tier using the annotating tool ELAN. We also offered them a list with predefined labels of emotions that they might see in the recordings. This list consists of 12 template facial expressions and the corresponding features. To reduce the chance on confusion we include a picture of each template on the list and description of it. The results are an annotated corpus containing the different facial expressions observed by annotators during the recordings.

4.2. Data Analysis

After annotation of the recordings we start analyzing the data. After some visual inspections of the annotated data it turned out that most of the labels given by different annotators were given to almost the same segments although they didn't agree on the given label all the time. They agree most of the time about the reflexes on the face which had no meaning. They could clearly distinguish between facial expressions and reflexes which don't show any emotional state.

Since we wanted to find out if emotions are perceived differently by different individuals, we proceed further with the experiment. We asked 3 different annotators to annotate our data to compare their perceptions. Two of them annotated the data together and were able to discuss the labels during the annotation with each other. The third one had almost no consultation with the rest but she had access to the annotated corpus by two other annotators. In appendix D an overview of all the assigned labels by annotators is given. For comparing the annotated corpus with each other we used inter-labeler agreement (Reidsma D., 2008). The inter-labeler agreement gives information about the level of agreement between the annotators and also suitability of the used labels and the difficulty of the labeling task.

To find the level of agreement between the annotators (inter-labeler agreement) the chance-corrected reliability metrics was calculated (Reidsma D., 2008) (Steidl S., 2008). It normalizes the levels of agreement from different annotation tasks with respect to chance-expected agreement for the task and makes them comparable with each other and expresses a chance-corrected level of agreement. It was introduced for the first time by Carletta in 1996 in computational linguistic (Krippendorff K., 1980) and is defined as follow:

$$k = \frac{P(A) - P(E)}{1 - P(E)}$$

Where P(A) is observed agreement among annotators and P(E) is agreement expected by chance.

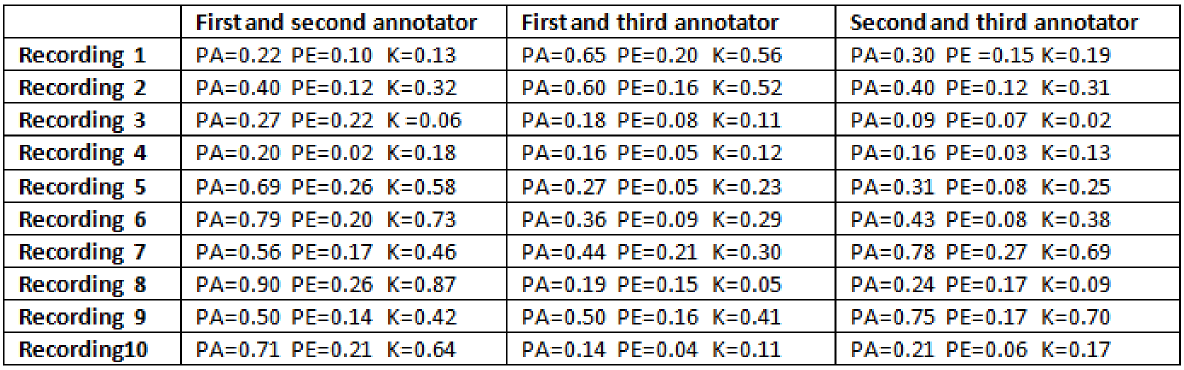
When the achieved agreement is exactly the same as what would be expected by the chance then k=0. When achieved agreement is perfect then k=1. A value of k =0.5 can be interpreted as the level of agreement for this annotation is exactly midway between perfect agreement and the level that would be expected by chance (Reidsma D., 2008).

A naïve way to calculate the level of agreement is to count the number of instances where two annotators agreed on the assigned labels compared to total number of instances the annotators had judged. In our annotated recordings (Appendix B) the first annotated recording contains in total, 13 assigned labels, by two of our annotators. From 13 assigned labels they agreed on 6 of them. The level of agreement here is 6/13 (0,46).

In the case of assigning the labels blindly and place them randomly without actually looking at the content of selected segment, the level of agreement reaches the value of 1/13 x 1/13 = 0.0059, which is very low and is not acceptable. No matter which label the first annotator chooses there is 0,07 chance for the second annotator to select the same label.

Using the chance corrected reliability metrics for comparing, the level of agreement obtained by our annotators is shown in table 5 (Kappa). The lowest level of agreement of table 5 is 0.02 and is reached while comparing the annotation of recording 3, done by second and third annotator. However, the reached level of agreement between first and second annotator in this recording is also not a high value. It shows the difficulty level of this task. The recording contains a lot of vague expressions and the suggested labels are not really matching the shown expressions in this recording.

Table 5: Results of calculating the inter-label agreements.



	First and second annotator	First and third annotator	Second and third annotator
Recording 1	PA=0.22 PE=0.10 K=0.13	PA=0.65 PE=0.20 K=0.56	PA=0.30 PE =0.15 K=0.19
Recording 2	PA=0.40 PE=0.12 K=0.32	PA=0.60 PE=0.16 K=0.52	PA=0.40 PE=0.12 K=0.31
Recording 3	PA=0.27 PE=0.22 K =0.06	PA=0.18 PE=0.08 K=0.11	PA=0.09 PE=0.07 K=0.02
Recording 4	PA=0.20 PE=0.02 K=0.18	PA=0.16 PE=0.05 K=0.12	PA=0.16 PE=0.03 K=0.13
Recording 5	PA=0.69 PE=0.26 K=0.58	PA=0.27 PE=0.05 K=0.23	PA=0.31 PE=0.08 K=0.25
Recording 6	PA=0.79 PE=0.20 K=0.73	PA=0.36 PE=0.09 K=0.29	PA=0.43 PE=0.08 K=0.38
Recording 7	PA=0.56 PE=0.17 K=0.46	PA=0.44 PE=0.21 K=0.30	PA=0.78 PE=0.27 K=0.69
Recording 8	PA=0.90 PE=0.26 K=0.87	PA=0.19 PE=0.15 K=0.05	PA=0.24 PE=0.17 K=0.09
Recording 9	PA=0.50 PE=0.14 K=0.42	PA=0.50 PE=0.16 K=0.41	PA=0.75 PE=0.17 K=0.70
Recording10	PA=0.71 PE=0.21 K=0.64	PA=0.14 PE=0.04 K=0.11	PA=0.21 PE=0.06 K=0.17

The highest value of table 5 belongs to annotating recording 8 by first and second annotator with k=0.87. This high value means that annotating this recording and comparing to other was not a difficult task. The first and second annotators agreed very frequently with each other about the shown expressions and the suggested labels. Since we see that the gained agreement values of this recording amongst other annotators are not really high. Many uncertainties between the first and second annotators were solved as they could consult each other. Since the third annotator had no consultation with the rest, the uncertainties caused a lot of disagreement between the given labels and caused a low level of agreement.

Let’s assume that the annotators are not annotating blindly and they do look at the content. Every disagreement between annotators may not have the same impact (Reidsma D., 2008). For example assigning the label ‘happiness’ by one of the annotators to the number of frames which is labeled with ‘sadness’ by another annotator is not the same as labeling a number of frames with ‘grief’ and ‘sadness’. In both cases the annotators don’t agree with each other but in the first case the distance between two labels is more than the second case.

Looking through the given labels we understand that most of the disagreements between the annotators are coming from categories that are not very far away from each other. A good example is the three labels ‘astonishment’, ‘surprise’ and ‘disbelief’. These 3 labels are very often given to the same number of frames by different annotators. After some more visual inspections of labeled data we understand because of using the non-speaking recordings in our experiment, lot of observed emotions in the recordings didn’t reach their real apex, and as consequence the differences between some of the emotions like ‘astonishment’ and ‘surprise’ are less visible. We decided to calculate the level of disagreement for the second time, considering that some of the labels belong to the same family and the differences between these labels is not the same as differences between labels from 2 various families. According to the theory of Ekman, human facial expressions are categorized in six basic emotions ‘happiness’, ‘sadness’, ‘anger’, ‘fear’, ‘disgust’ and ‘surprise’. These 6 emotions are innate and universal across different cultures. The other facial expressions are culture dependent and learned during life. They form the families of the basic emotions. Each member of an emotion family shares certain characteristics with other member which makes them belong to the same family. To calculate the level of agreement for the second time, we categorized our labels in 6 basic emotions. Emotions like ‘sad’, ‘regret’ ‘grief’ form one label. After categorizing the labels the number of labels is reduced from 12 to 5. The new categorized labels are displayed in the table 5.

Table 6: Categorized labels.

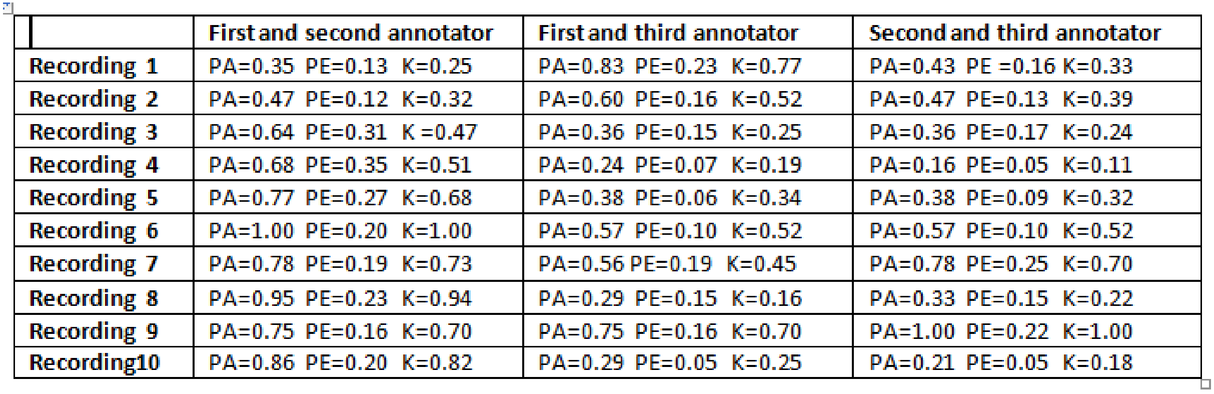
Happy	Sad	Anger	Fear	Disgust	Surprise
Happy	Sad	Anger	_______	Disgust	Surprise
Satisfaction	Regret	_______	_______	_______	Astonishment
Ironic Smile	Grief	______	_______	_______	Disbelief
_______	_______	_______	_______	_______	Understanding

During the experiment we asked the annotators to assign a label to the parts of recordings when they observe a facial expression. But the labels were not always necessary assigned to the exactly same frames and start and end of the selected frames differ with each other. It makes defining the value of agreement more complicated. For solving this problem we defined a new value, the θ value. Using a θ value makes the comparisons of the labels that don’t start and end on exactly same frame possible. The start and end time of the selected frames should not differ with each other more than θ value. The θ value depends also on the length of the labels, very long labels should overlap in more frame comparing with the short segments. Short segments need to conform to a smaller threshold differences in timing (Reidsma D., 2008).

Looking through our data we observed that the elapsed time for most of selected facial expressions is not longer than 1.5 seconds (37.5 frames) and the longest one is about 3 seconds. To determine a good level of agreement we start estimating our θ value. For the segments shorter than 1.3 seconds we decide θ = 0.3s and it means the segments shorter than 1.3 seconds may not differ with each other in start and end time in more than 0.3s. For the segments longer than 1.3 the θ= 0.5s. These labels cannot differ in start and ending frames more than 0.5 second and if they differ in start and end more than 0.5 seconds they are not considered as the labels assign to the same segments.

The results of recalculating the inter-labeler agreement levels after considering the θ value and the fact that some of the labels are part of the same family are given in table 7. As you can see the levels of agreement are higher than the levels of agreement from table 5. After categorizing the labels the highest level of agreements reaches k=1.00 which is the highest possible value for a level of agreement. This reached value for some recordings shows that the annotators didn’t make a mistake performing their task, the offered labels perfectly match the observed facial expressions and the recordings don’t contain very vague expressions. The lowest value of this table is k=0.11 which belongs to annotating recording 4 done by second and third annotator. It may be noted however, that most of the values of 'k' are much below 1.00 which means that the annotators don't agree a lot with each other. This could depend on experience and background of annotators.

Table 7: Results of recalculating the inter-label agreements.

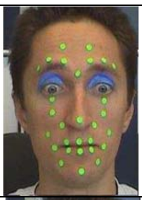
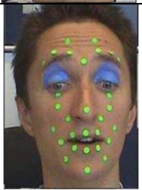
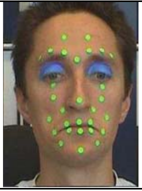
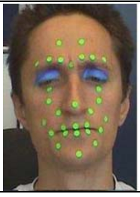


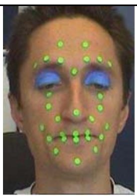
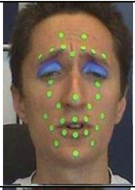
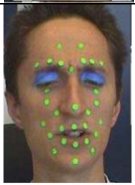
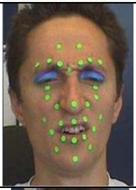
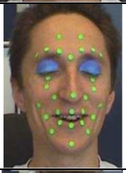
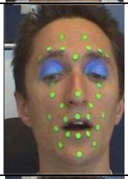
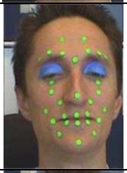
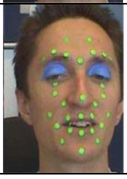
	First and second annotator	First and third annotator	Second and third annotator
Recording 1	PA=0.35 PE=0.13 K=0.25	PA=0.83 PE=0.23 K=0.77	PA=0.43 PE =0.16 K=0.33
Recording 2	PA=0.47 PE=0.12 K=0.32	PA=0.60 PE=0.16 K=0.52	PA=0.47 PE=0.13 K=0.39
Recording 3	PA=0.64 PE=0.31 K =0.47	PA=0.36 PE=0.15 K=0.25	PA=0.36 PE=0.17 K=0.24
Recording 4	PA=0.68 PE=0.35 K=0.51	PA=0.24 PE=0.07 K=0.19	PA=0.16 PE=0.05 K=0.11
Recording 5	PA=0.77 PE=0.27 K=0.68	PA=0.38 PE=0.06 K=0.34	PA=0.38 PE=0.09 K=0.32
Recording 6	PA=1.00 PE=0.20 K=1.00	PA=0.57 PE=0.10 K=0.52	PA=0.57 PE=0.10 K=0.52
Recording 7	PA=0.78 PE=0.19 K=0.73	PA=0.56 PE=0.19 K=0.45	PA=0.78 PE=0.25 K=0.70
Recording 8	PA=0.95 PE=0.23 K=0.94	PA=0.29 PE=0.15 K=0.16	PA=0.33 PE=0.15 K=0.22
Recording 9	PA=0.75 PE=0.16 K=0.70	PA=0.75 PE=0.16 K=0.70	PA=1.00 PE=0.22 K=1.00
Recording10	PA=0.86 PE=0.20 K=0.82	PA=0.29 PE=0.05 K=0.25	PA=0.21 PE=0.05 K=0.18

4.3. The offered labels

The annotators were asked to find and mark 12 pre-defined labels in the recordings. These labels contain 12 emotional states defined by Anna Wojdel and were selected by visual inspection of the recordings and looking for the easily classifiable types of facial deformations. These facial deformations were selected independent from deformations resulting from the speech. For selecting the template facial expressions only the distinct facial deformations were considered and eventually the 12 template facial expressions were distinguished (Wojdel A., 2005). The 12 template facial expressions and the features are shown in table 8.

Table 8: Template expressions.

No	Label	Features	Template expression
1	astonishment	raised eye-brows and eyes wide open	
2	surprise	raised eye-brows	
3	sadness	lowered corner of the mouth, raised chin	
4	disbelief	lower eye-brows and mouth slightly stretched	

5	regret	tightened and stretched mouth	
6	grief	raised inner eye-brows	
7	anger	lowered eye-brows	
8	disgust	wrinkled nose	
9	happiness	open mouth, raised corner of mouth and raised cheeks	
10	understanding	withdrawn and lifted up head, mouth open and slightly raised eye-brows	
11	satisfaction	slightly raised chin and corners of the mouth	
12	ironic smile	raised upper lip and corner of the mouth	

Labels assigned to each template expressions in table 8 were chosen on the basis of appearance of facial features, independently from the context in which they occur. They were selected to easily refer to specific expressions and therefore their real meaning in the recordings is not always adequate to the label (Wojdel A., 2005).

4.4.Triggers

Analyzing text and recognizing the emotional words is crucial for understanding of the background information in that text. The question is which part of the text triggers an emotion. We observed the annotated video sequences and look to the translated text. We used the translated text because the recordings were made in Polish and as the text has a big contribution to our interpretation we provide the translation of the scenarios to annotators. So it can be expected a displayed emotion can be explained from the context. Analyzing the text and observed emotions we found the following triggers:

1. Emotional words.
2. Semantic of a text.
3. Punctuation.

Emotional words in a text are the words which convey an effect immediately. The most used method for machine's textual affect sensing is key word spotting, or looking for emotional words. In this method text is classified into effect classes based on the presence of effect words like 'happy' or 'angry'. Most of the times an emotional word triggers an emotion. Emotional word 'happy' triggers most of the time a positive effect in us. Weakness of this approach is poor recognition of effect when negation is involved. Understanding the semantic of a text enables us to not only learn the effective valence of effect key words but also take into account the valence of other arbitrary keywords. Paying attention to the semantic of a text helps us to understand the real meaning of the sentence. For example the sentence 'I believe that you are not happy' despite the emotional word 'happy' doesn't contain a positive valence and keyword spotting doesn't provide us the correct background information of the text to trigger the right emotion.

Punctuation in a text also has a big impact on extracted emotions from a text. In fact in some cases they can change the meaning of a sentence. In the following example the meaning of the sentence is strongly influenced by the punctuation.

A woman, without her man, is nothing.
A woman, without her, man is nothing.

This example indicates the importance of the punctuations. In daily life, a sentence ending with a question mark easily extracts 'uncertainty' or an exclamation mark most of time extracts 'astonishment'. We got the same results from our annotated corpus. After analyzing our annotated corpus we noticed the big role that punctuations play to trigger the emotions. They help to sense the effect from text which is a big contribution for understanding the observed emotion. On the other hand we should not forget the influence that punctuation has on displayed emotions by the actors during the recordings. Ending the sentence 'you were at her place tonight?' with a *question mark* extracts 'astonishment' as in the actor during the recording and as in the annotators during annotation.

4.5. Duration and frequency

In our attempt to analyze the labeled facial expressions we checked whether facial expressions differ in respect to the number of occurrences and their lengths. We analyzed the annotated facial expressions of recordings of a listener done by Anna Wojdel (as explained in part 3.2.) and the results may be slightly different from the results of analyzing the recording of a dialog. In these recordings the most common facial expression is 'satisfaction' and the least appearing labels in our recordings are 'regret', 'disgust' and 'sadness'.

The recordings last about 13.19 minutes and they contain in total 19975 frames. Annotated segments done by first annotator contain 3048 marked frames and the segments annotated by second annotator contain 3072 marked frames. Almost 15% of the frames are marked by annotators as frames displaying facial expressions and the rest are considered as neutral emotional state. In table 9 the statistics of manually selected facial expressions by two of our annotators are shown. The longest average of the labels belong to 'happiness' and the shortest average belongs to 'sadness'. 'Satisfaction' is the most frequent label with the average length of almost 31 frames which is about 1.4 seconds.

Table 9: Statics of manually selected facial expressions by two annotators.

	(annotator 1)				**(annotator 2)**			
Label	**No. of segments**	**No. of frames**	**% of annotated frames**	**Average no. of frames**	**No. of segments**	**No. of frames**	**% of annotated frames**	**Average no. of frames**
Disbelief	17	600	20	33.6	19	624	22	32
Surprise	13	384	13	29.5	10	312	12	31.2
Satisfaction	28	1056	35	37.7	22	888	28	40
Happiness	4	168	5	42	3	168	5	56
Understanding	8	216	7	27	7	240	7	34.2
Anger	7	168	5	24	3	120	3	40
Grief	1	24	0,7	24	2	48	1	24
Ironic Smile	5	192	6	38.4	12	408	13	34
Regret	1	24	0,7	24	0	0	0	0
Astonishment	5	168	5	33.6	7	192	6	27.4
Disgust	1	24	0,7	24	2	48	1	24
Sadness	1	24	0,7	24	1	24	0.5	24

Comparing the minimal and the maximal length of the segments we notice that the most of the facial expressions are expressed for a very short period of time, 1 or 1.5 seconds. As these are results of analysis of annotated listener recordings of Anna and because of absence of talking mouth in the recordings, the results would be slightly different from the results gained in analyzing the recordings of a dialog. In this annotated corpus the facial expressions are expressed in a longer time than in a dialog. We usually show a part of our emotions also through talking, when we are showing our emotions just through facial expressions and not through speech we see stronger facial expressions in some part of face like in eyebrows and eyes which may last for a longer period of time and less facial expressions in some other part of face like in contour of mouth. In the recording of a dialog the most of the facial expressions are expressed between a 0.5 of second and 1 second (Wojdel A., 2005). In recordings of listener, the average is between 1 and 1.5 seconds.

In figures 14 and 15 the numbers of assigned labels to the different segments by two annotators are shown. As we said the most frequent facial expression is 'satisfaction' and the average length of the segments is 1.5 frames. After 'satisfaction' the most common expressions are 'disbelief', 'ironic smile' and 'surprise'. The lowest number of occurrence belongs to 'grief', 'sadness', 'regret' and 'disgust'.

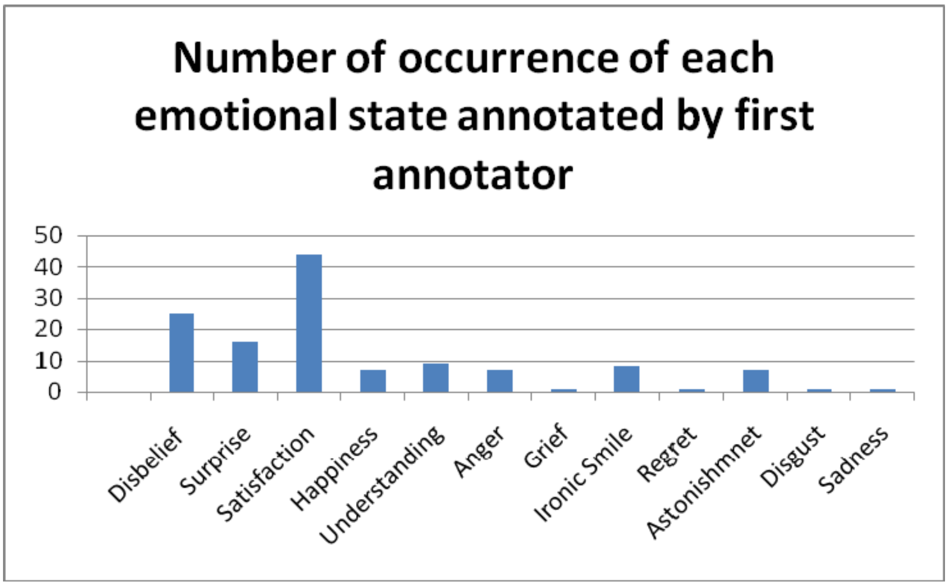


Figure 15: Number of occurrences of different emotional states done by first annotator.

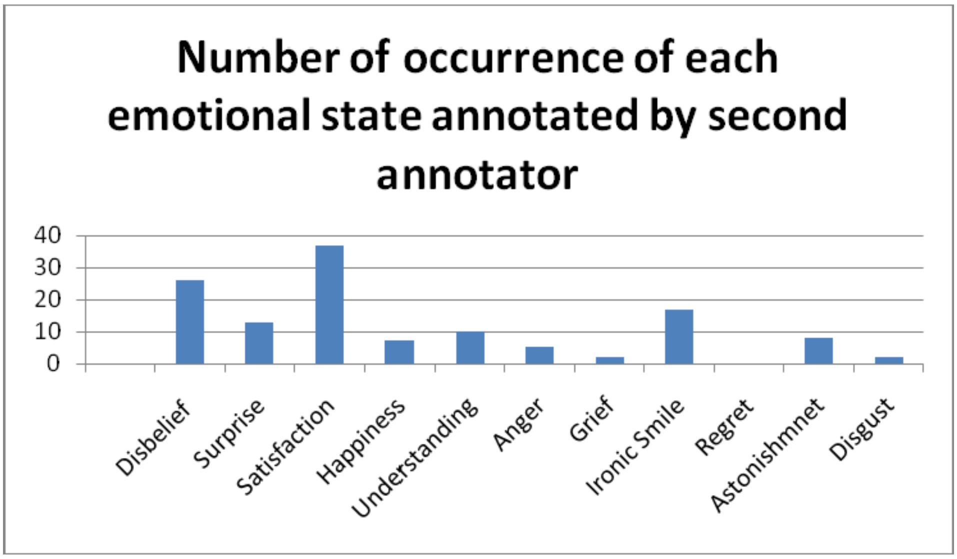


Figure 16: Number of occurrences of different emotional states done by second annotator.

Chapter 5

5. Model of tracking points

Facial expressions are generated by activation of facial muscles. The activation of muscles is not directly visible but by changes of contour of eyes, eyebrows, mouth and etc. Many methods are designed for tracking the changes in facial expressions. But these methods are limited by problems such as lighting conditions, postures and occlusions. To facilitate the tracking process we put markers on the face on places of supposed muscle's activations. Now we have to localize the markers and track the markers in the video recordings to find cluster of facial emotions.

Human observers have no problems in localizing and tracking the green markers. But an automated tracking tool is far from trivial. The green markers will have different shapes and colors in changing light and posture conditions. The light condition isn't same for all viewing angles and markers on the face have a different shape from different view. We took video recordings from the facial expressions at a rate of 24 frames per seconds. But some facial expressions change at a higher rate. So to track the green dots a special procedure is needed as explained in this chapter and we describe the tracking of the markers from the face-to-face communication recordings. The locations of localized markers are identified on each frame and are saved in the vectors. The facial features for each emotion after analyzing the movements of tracked points are defined and these facial features can be used for building fully automatic facial feature expression recognition from face-to-face communication.

5.1.Prototype

During our experiment we created an annotated corpus (as explained in chapter 4) and we cut the recordings in the annotated segments with the assigned labels. Using our model we load an annotated segment (sequence of images) and the processing starts. The first step of processing is extraction of the frames from the video clip. Then our model tracks the face and facial features from each frame. The feature points on the face per frame are recognized and saved to a vector. In this part we explain how the model processes each frame, tracks the points and saves the movements.

After extracting the frame from video recording the face is cut out in Matlab using the Machine Perception Toolbox. We convert the image of the face with points on it to HSV and extract red and green components. To extract the green dots and the blue blobs we set some threshold values. For each frame we start with a fixed color based filter to the pseudo hue space of the image. We prefer working on the pseudo hue based color space because the pseudo hue color segmentation gives better results than HSV color segmentation. We calculate the pseudo hue component as follow:

$$Ph = \frac{R}{R+G}$$

Where P_h is the *pseudo hue* component, R is the *red* component and G is the *green* component of the image. Then we segment the image and extract the part corresponding to the green stickers.

$$F(x,y) \rightarrow \begin{cases} 1, & P_h > 0.11 \\ 1, & v < 0.5 \\ 0, & \text{Otherwise} \end{cases}$$

Where F(x,y) is the value of the binary image on location (x,y) and *v* is the *value* (HSV) component of the image. An example of extracted green points is shown in the image 17. In the following piece of Matlab code the image returned by the face detector is processed in order to detect the green dots and the blue eyes. This method as we said before makes use of the pseudo hue color space. It uses the color image from face detector and produces the binary image of green dots and blue eyes.

```
img_face_hsv = rgb2hsv(img_face);           % convert image to HSV

h = img_face_hsv(:,:,1);                    % segment the image
v = img_face_hsv(:,:,3);

red = double(img_face(:,:,1));              % extract red and green components
green = double(img_face(:,:,2));

ph_img = imdivide(red,imadd(green,red));    % calculate pseudo hue component

ph_img = histeq(ph_img);                    % perform histogram equalization

img_face_out = (ph_img<.11) & (v>.5);       % extract only the corresponding
 …                                          % pseudo hue values
 …

eyes_original = (h>.6) & ( h <.9) & (v>.5); % get roughly the eyes
```

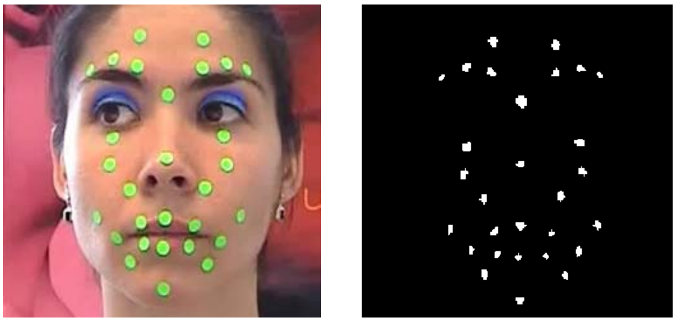


Figure 17: The original image and the detected green points on the face by processing.

Extracting the blue blobs is almost done in the same way. Where F(x,y) is the value of the binary image on location (x,y) and *h* is the *hue* (HSV) component of the image.

$$F(x,y) \rightarrow \begin{cases} 1, & h > 0.60 \\ 1, & 0.5 < v < 0.9 \\ 0, & \text{Otherwise} \end{cases}$$

An example of the extracted blue eyes is shown in the image 18.

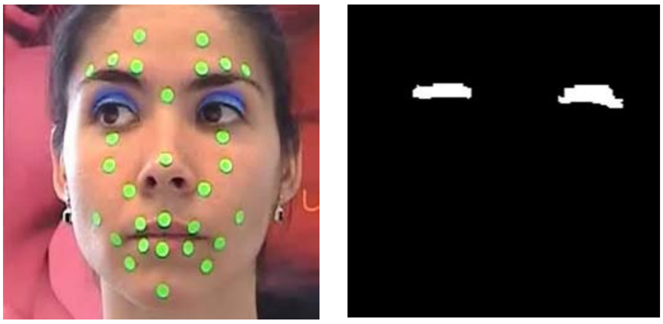


Figure 18: The original image of the face and the detected blue blobs by processing.

Using the green points and blue blobs are just for easier detection of points. The blue eye shadows are to avoid sticking a sticker on eyelid of the participant. In the future it is possible to adjust the system to accept facial points found by other algorithms than the color segmentation.

5.2.Labeling and validation

After recognizing 29 green points and 2 blue blobs they should get labels. We start with labeling 5 key points on face. Our key points on the face are: the lowest point on the chin, the point between the eyes (the closest point to the center of mass of the labeled eyes), the point on the nose and the points on the upper lip and the lower lip. The first point to label is chin points. To label this point first the image of the face is -90 grades rotated. At this moment the chin point should be the first point from the left side of the image. After labeling the chin point, image is 90 grade rotated back to original position. These steps are shown in figure19. Then we start labeling the rest of the key points. To do this we use the vertical and horizontal dependencies between the points. We explain more about the vertical and horizontal dependencies in the following page.

After labeling all the key points to see if the head was rolled during recording, the angle of face is calculated using the angle between chin point and eye point. The face is rotated to the position that the line between chin and eye points is perpendicular to the horizon line. After Face Straightening it is time to normalize the distance between two key points, namely eye point and nose point. To normalize the distance between these points the distance between the eye point and the chin point is set to 50 pixels. This way it doesn’t matter how big the image is or the face is moving forward or backward.

```
                                      % label the different dots
greenDots = imrotate(greenDots,-90);  % Rotate the image -90 grade
greenDots = bwlabeln(greenDots);      % if correct there must be 31 stickers visible
greenDots = imrotate(greenDots,90);   % First point is now the chin-point
```

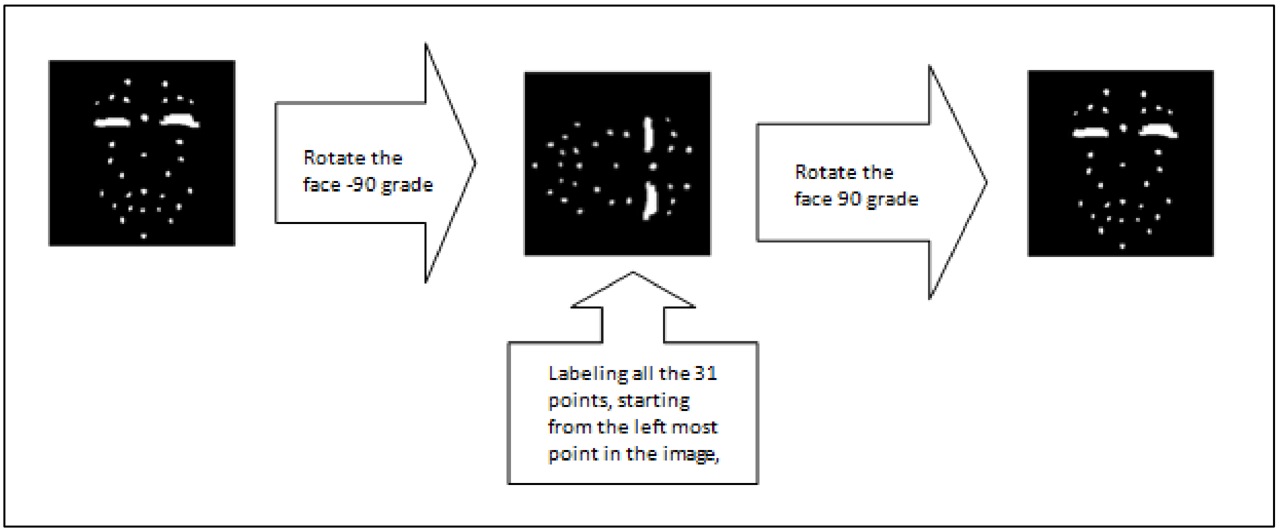

Rotate the face -90 grade
Rotate the face 90 grade
Labeling all the 31 points, starting from the left most point in the image,

Figure 19: Flowchart of labeling process.

After labeling all the points on the face the labeled points are validated to see if they are recognized and labeled correctly. If some of the points are not recognized well or are not recognized at all they will be recalculated based on vertical and horizontal dependencies between some of the points. These dependencies are shown in figure 20. We assume that the key points (points between the eyes, point on the noise, on the upper and the lower lip and on the chin) are always well recognized and based on the place of these points the place of other points can also be calculated. These dependencies between the points and the symmetry of the face help to find the right positions of the points that are miscalculated or the points that are not found at all. When all the 31 points are well recognized we start saving them in a vector frame by frame for each annotated segment. Saving the points is done real time. Per frame recognized points are saved to a vector.

Anna didn't use diffused light during the recordings and it caused a lot of lighting problems like shadows on the face. Because of mentioned problem not all the green dots were recognized well using the tracking algorithm and the shape of found points were not always round. The consequences are the centers of the points which indicate the final location of the points were not on the same place on each frame even when there were no real changes in the location of the point. Because of this incomplete shape recognition we don't count the movements smaller than 0.5 pixels as the movements indicating changes in the positions of the points on the face and they are assumed as errors of the program.

After finding the points, if there are still some points missing, we recalculate them using the positions of the neighboring points. Most of the points have vertical or horizontal dependencies with other points and using these dependencies we recalculate the placement of the missing point. Also the symmetry of face can help us to recalculate the missing points if at least one of the mirroring points is found. If both of the mirroring points are missing using the previous location of these points can help to reconstruct the present location. At the figure 20 some of the horizontal and vertical dependencies between the points are shown. Horizontal dependencies are shown with blue lines and the vertical dependencies are shown with pink lines.

During the processing part we found out that recognizing of all the 31 points without using the vertical and horizontal dependencies between the points doesn't happened very often. Due to the

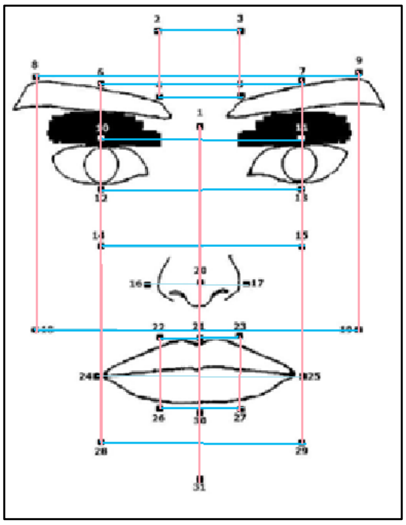


Figure 20: Dependencies between points.

lighting problem there are only few frames where all the 31 points are recognized at first place without the recalculating algorithm. Using the dependencies all the points are eventually found and saved in a vector.

Chapter 6

6. Results of tracking

After some visual inspection of data we noticed that some of the points on face show more movement than others. For example the points on the eyebrows and the points around the mouth are moving more often than the points on the forehead or the points on the cheeks. In this chapter we follow some points on the face to extract facial features for each emotion and define different emotion clusters. Using the extracted features helps us to make an algorithm for recognizing the emotion. Based on the extracted features and the found cluster, we assign a label to the recording. An example is shown in figure 21. Following the movements in this example we reached cluster 'surprise' and in the next frames we ended up in cluster 'happiness'. If the extracted features guide

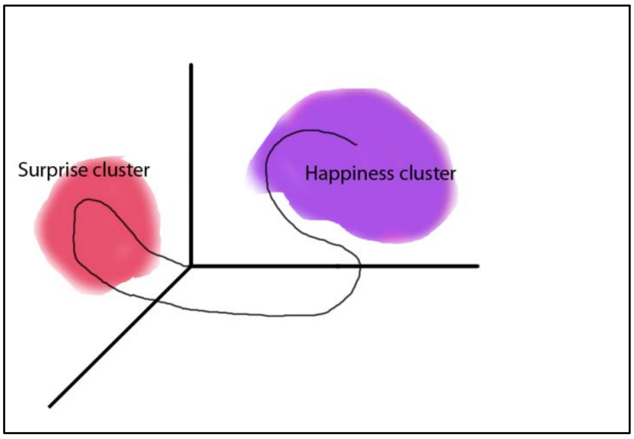


Figure 21: Following the movements on the face to find the emotion clusters.

us in an area between two or more clusters the closest distance to one of those areas provides us the right label.

6.1.Tracking of selected points

After tracking the points on the face and saving them in a vector, we chose two crucial points among all the points on the face to follow. After some visual inspections we decided to divide the face in two areas, upper side area and lower side area. Among all the points in the upper area the points on the inner side of the eyebrows show the most movements during expression of emotions. The movements of these points on the eyebrows were most visible. We chose to follow the point on the inner side of right eyebrow for each expression. The number assigned to this point in our database was 4 and we refer to this point from now on as point 4. In figure 22 this point is indicated with a red marker.

The second point we chose to follow was a point from lower side of the face. To find the right point from this area we did some visual inspections through our database. The chosen points from the lower area were the points on the mouth corners. These points indicate most movement of the lips. We selected the point on the right corner of the mouth, as it showed enough movements comparing to other points. The number assigned to this point is number 24 and from now on we refer to this point as point 24. In figure 22 this point is marked with blue.

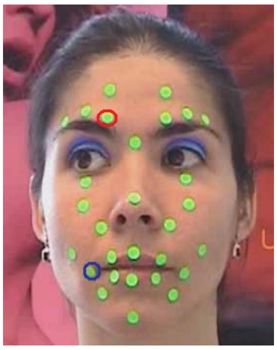


Figure 22: Point 4 and point 24.

To observe the movements on the face for extracting the features and define the emotion clusters we chose a part of a recording with some transitions between emotions. In this example we follow point 4 starting from a natural position and follow it during a transition to the emotional state 'disbelief'. The green path shows the cluster 'disbelief' and the path that point 4 during this transition followed. As it is to seen in figure 23 during emotion 'disbelief' this point goes low. The part of the path in blue shows the emotional state of 'grief' and the supposed cluster of 'grief'. During this emotional state the point 4 goes up. Finally when the face goes back to natural position the point goes almost back to the place it started.

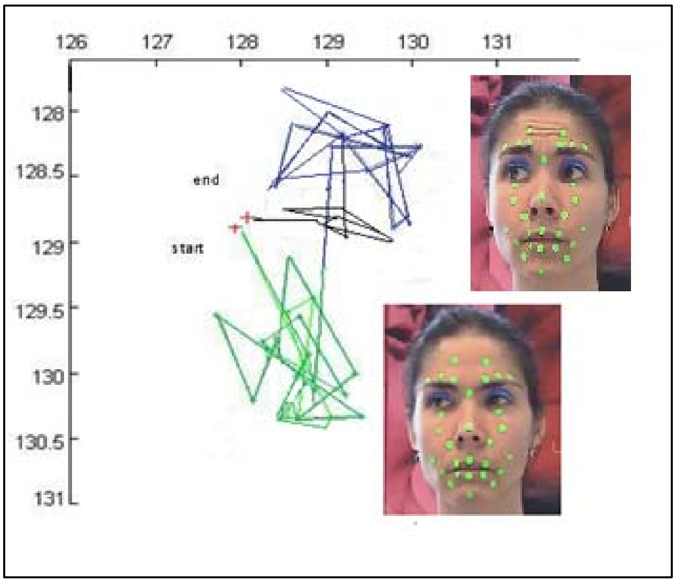


Figure 23: Following point 4 in a recording containing a transition from 'disbelief' to 'grief'.

Following the movements of the points on the face we extract lots of features which enable us to define different emotion clusters. Each emotion cluster has almost the same facial features. Having more features for each cluster makes it easier to find a correct label for each recording.

6.1.1. Points on the eyebrows

The expected movements in some points are more visible than the others. As we explained in previous part we followed the points which are less dependent to the fact that we are using the listener recordings. In figures 24-26 we followed the movement of point 4 (on the inner side of right eyebrow) and point 5 (on the inner side of left eyebrow) during the emotional state of ’surprise’ to extract the concerning features. During this recording the eyebrows are moving upwards and after reaching the apex frame, they both come to the former place. At image25 and 26 the distances between different frames and apex frame (frame with full emotion) are shown. The distances at the first frames are bigger and as we approaching the apex frame the distances are getting smaller until at the apex frame the distance is zero. After passing the apex frame the distance is getting bigger till it almost reaches the original distance. The small movements (less than 0.5 pixels) are assumed to be errors of the program and they don’t show the real movements. In the figures 24 point 4 is marked with red and point 5 is marked with blue. In this figure we chose 3 different frames from recording ‘surprise’. First frame is chosen from the beginning of the recording, second frame is the apex frame and shows the full emotion and the last frame is one of the last frames of the recording. In figure 25 we followed the movements of point 4 and plot the distance of each frame with the apex frame. The red marked points correspond with the frames from figure 24. In figure26 point 5 is followed and the distance of each frame from apex frame is plotted. The blue marks on the plot correspond with the chosen frames of figure 24. Because the used recordings are listener recordings we observe almost no mouth movements.

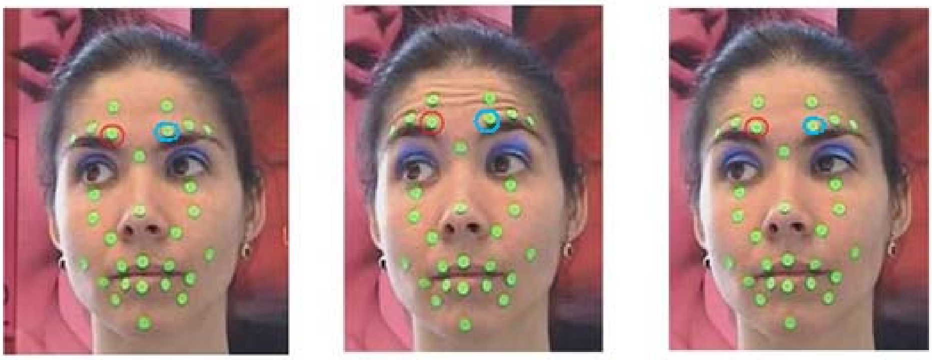


Figure 24: Movement of point4 during emotional state 'surprise'.

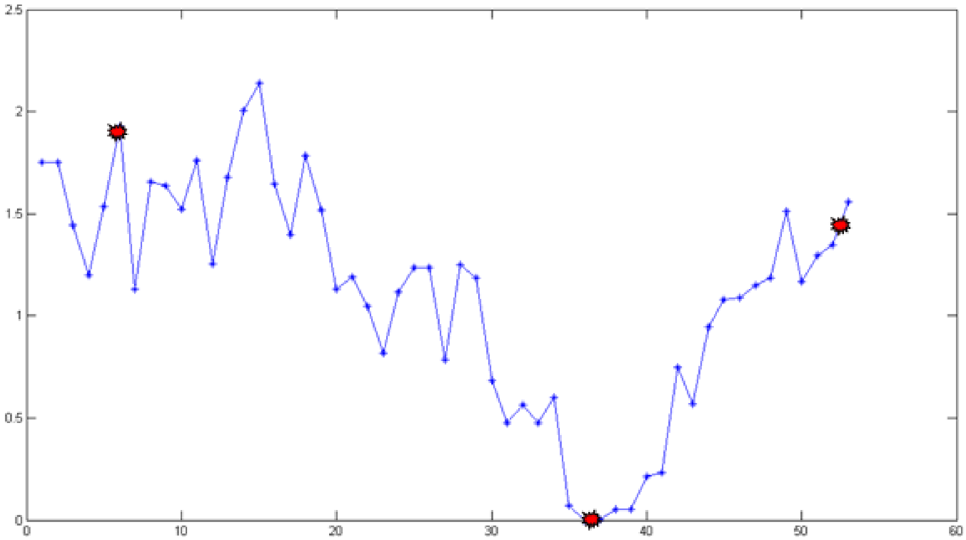


Figure 26: Distances between point 4 on apex frame and other frames.

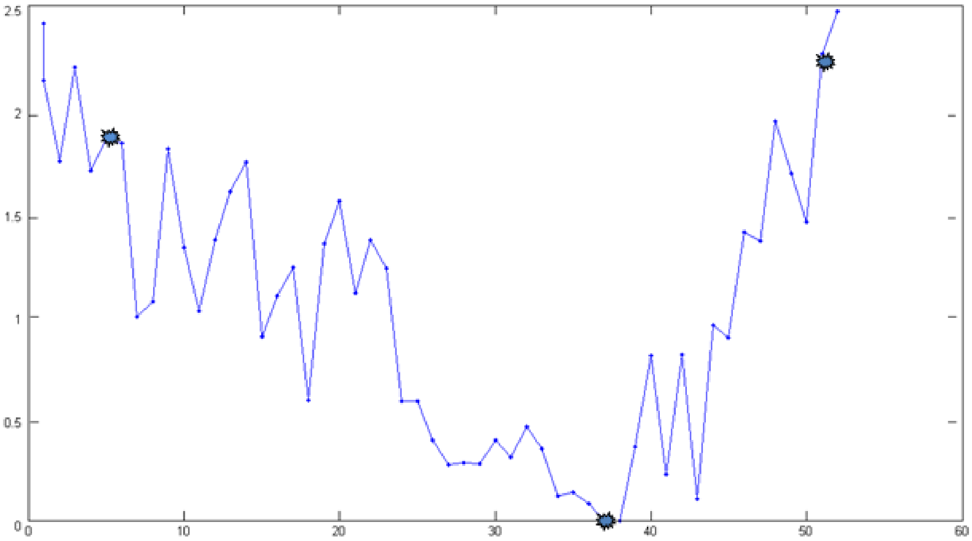


Figure 25: Distances between point 5 on apex frame and other frames.

6.1.2. Points around the mouth

In this part we followed two points from the lower part of the face to analyze their movements and extract features. The chosen points are the points on the corner of mouth during the emotion happiness.

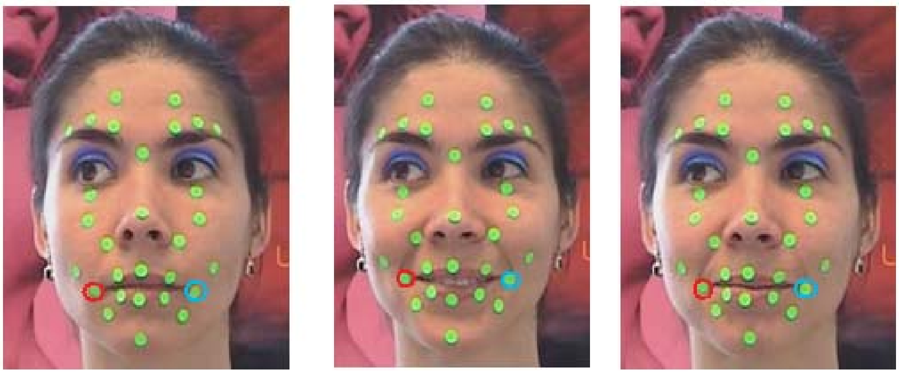


Figure 27: Movements of point 24 and 25 during 'happiness'.

Point 24 is the point at left corner of the lip and point 25 is the point on the right corner of the lip as it is shown in figure 27 with red and blue markers. During ‘happiness’ in the chosen recording we observe a smile which left corner of the mouth moves more to the left and right corner of the mouth moves more to the right. These movements are shown in the figure 28 and 29.

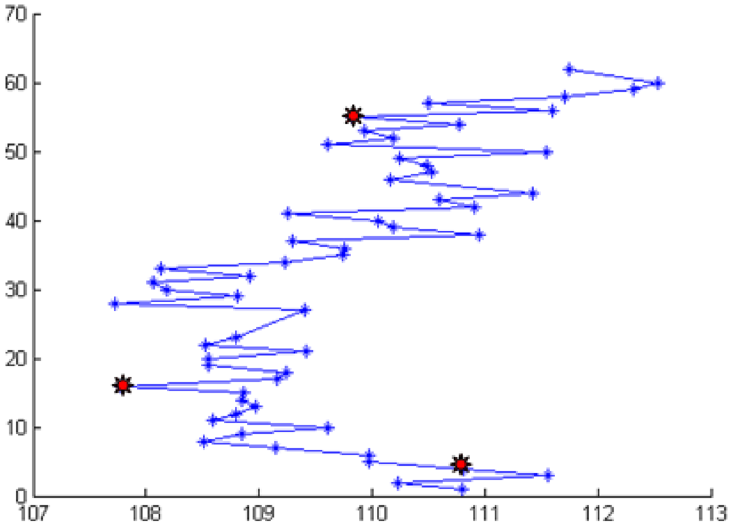


Figure 28: Following of the movements of point 24 during 'happiness'.

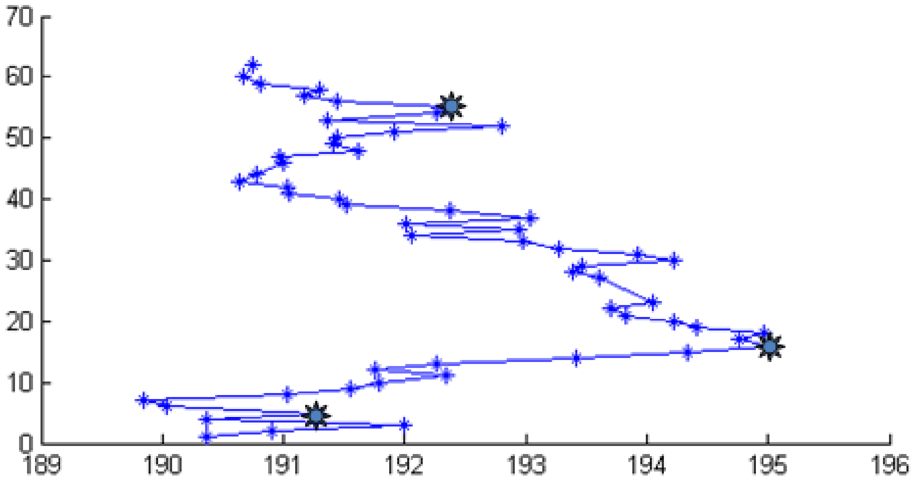


Figure 29: Following of the movements of point 25 during 'happiness'.

As we showed in the last two examples following the movements of some important points per emotion, helps us to extract facial features to define emotion clusters and recognize the different emotions from each other. Per emotion we should consider the movements of concerning points. For emotions like 'surprise' and 'anger' the movements of the points on eyebrows play an important role and for emotions like 'happiness' and 'sadness' the points around the mouth have more influence. As we can see from the plots, the tracking of data to extract features provides also lots of noise. But despite the noise these movements start from neutral position and slowly reach the apex and finally the point goes back to neutral position. Extracted features of each emotion expressed by different people provide us almost similar results. For extracting more features of all emotions we need more annotated samples of more people from different gender, age and color to define the emotion clusters more accurate.

6.2. Principal component analysis

In this part we start with an introduction to principal component analysis and then we explain why and how we used PCA on our data and finally we present the results. Principal Component analysis is a way of identifying patterns in data, and expressing the data in such a way to highlight their similarities and differences. Since in data of high dimension patterns can be hard to find and the luxury of graphical representation is not available, PCA is a powerful tool for analyzing data. The main advantage of PCA is that once the pattern is found in the data, the dimensions of the data are reduced without much loss of information (Shlens J., 2005).

Our data contains 10 different emotions (We started with 12 emotions but later we reduced it to 10, because we could not find enough samples for 2 emotions, 'disgust' and 'regret') and for each emotion we have almost 8 different samples and each sample consists of different numbers of frame. From each frame we tracked 31 points on the face and each point has 2 coordinates, x and y. That is 62 dimensions per frame per sample. To work with this high dimension data we applied PCA to our data. We did it in Matlab and made a small program to apply the PCA on our data and reduce the dimensions.

```
matrix=xlsread('LaatsteSerie_Apex.xls','B2:BR63');
[mn,mm]=size(matrix);
mat=matrix';
[mata,matb]=size(mat);
mu=mean(mat);
XX=mat-repmat(mu,mata,1);

c = cov(XX);                % find covariance matrix
[V,D] = eig(c);             % find eigenvectors (V) and eigenvalues (D) of
                            % covariance matrix
[D,idx] = sort(diag(D));    % sort eigenvalues in descending order by first
                             % diagonalising eigenvalue matrix, idx stores order
                             % to use when ordering eigenvectors
D = D(end:-1:1)';
V = V(:,idx(end:-1:1));       % put eigenvectors in order to correspond with
                              % eigenvalues
V2d=V(:,1:3);                 % (significant Principal Components we use,
                              % OutputSize is input variable)

prefinal=V2d'*XX';
final=prefinal';              % final is normalized data projected onto eigenspace
final = (-1)*final;
```

We apply the PCA to 31 points on the face of the apex frame per sample. We calculate the covariance matrix of the data, the Eigen value and the Eigen vector of it.

In the following figures the first two principal components of our data at three different moments of recordings are shown. Figure 30 shows the first and second principal components of points at some frame before apex. At figure 31 are the first and second principal components of the data exactly at the apex frame plotted. Figure 32 shows the first two principal components at some frames after apex. In figure 30 we observe less variation in data than in other plots. As all of our recordings start with a natural face and this plot shows the variation of the data at the starting frames we have less variation in data at figure 30 than in figure 31 and 32.

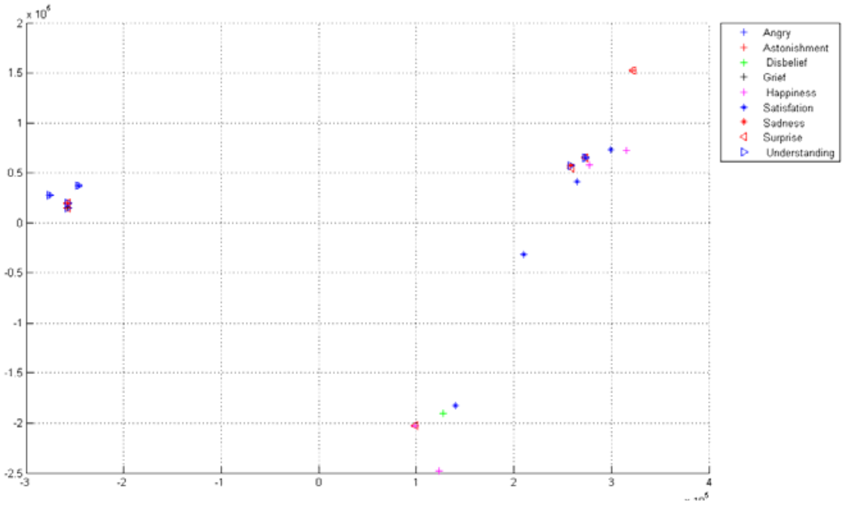


Figure 30: First and second principal component at some frames before apex frame.

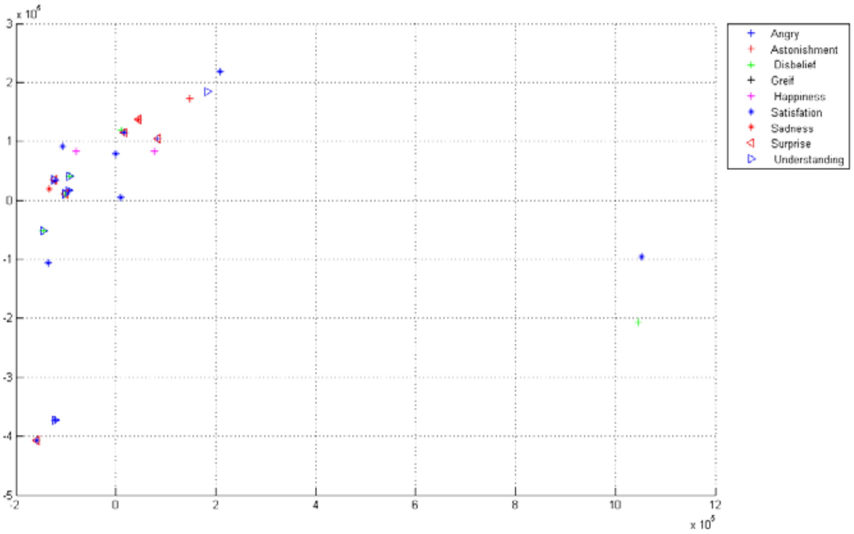


Figure 31: First and second principal component at apex frame.

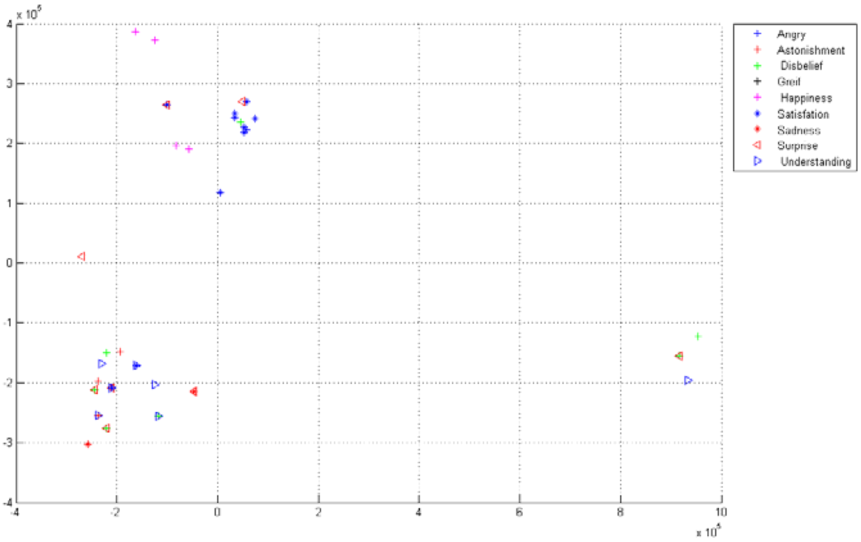


Figure 32: First and second principal component at some frame after apex frame.

Based on these plots we can't make a comment about our data and we can't find clusters. The first and second principal component of our data represent only 53% of the whole data which is not representative for the whole data and we lose almost half of our data.

6.3.Euclidean distance matrix

Because applying PCA and plotting the results in 2D didn't give us the expected results, we decided to calculate Euclidean distance between different emotions to compare the differences between them. Euclidean distance or Euclidean metric is the distance between two points (Gower J. C.). We start with Euclidean distance for 1d points and it is calculated as follow:

$$| p_x - q_x | = \sqrt{(p_x^2 - q_x^2)}$$

Where $P = (p_x)$ and $Q = (q_x)$. The absolute value signs are used since distance is normally considered to be unsigned scalar value.

Our data is not one dimensional and we need a formula for calculating the distance in 31 dimensional. The formula for N-dimensional space is given here:

$$P = (p_1, p_2, p_3, \ldots, p_n), Q = (q_1, q_2, q_3, \ldots, q_n)$$

The distance is computed as:

$$P - Q = \sqrt{((p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2)}$$

And for our data as we have a 31-dimensional data we use the formula as follow:

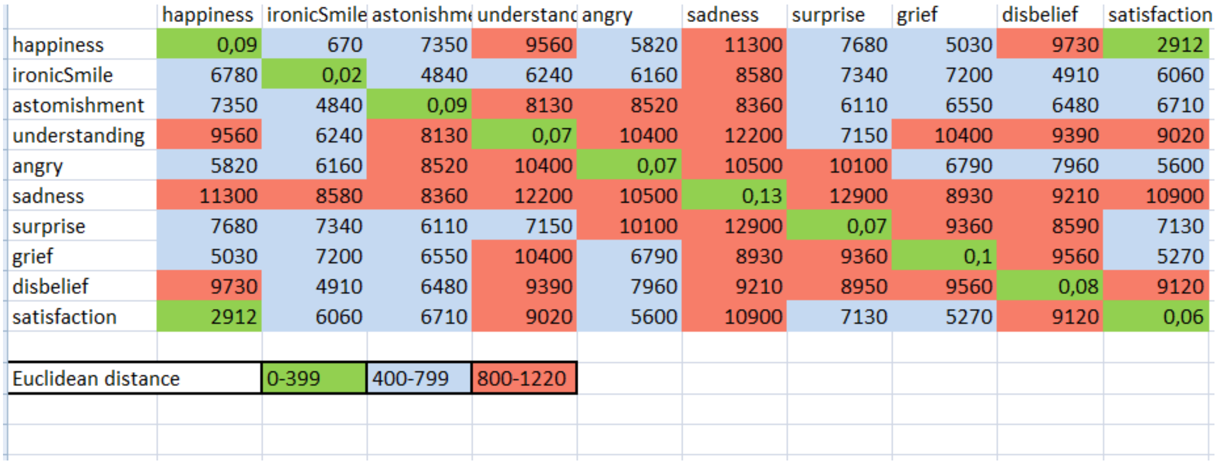
$$P = (p_1, p_2, p_3, \ldots, p_{31}), Q = (q_1, q_2, q_3, \ldots, q_{31})$$

The distance is computed as:

$$P - Q = \sqrt{((p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_{31} - q_{31})^2)}$$

The Euclidean distances between different emotions in our database are calculated using the last formula and given in the table 10. In this table we marked the values with 3 colors in 3 categories. Green cells show the distances between emotions smaller than 400 pixels. Red cells indicate the distances bigger than 799 pixels and blue cells indicate the distances between 400-799 pixels.

Table 10: Euclidean distances between different emotions from our recordings.



	happiness	ironicSmile	astonishme	understanc	angry	sadness	surprise	grief	disbelief	satisfaction
happiness	0,09	670	7350	9560	5820	11300	7680	5030	9730	2912
ironicSmile	6780	0,02	4840	6240	6160	8580	7340	7200	4910	6060
astomishment	7350	4840	0,09	8130	8520	8360	6110	6550	6480	6710
understanding	9560	6240	8130	0,07	10400	12200	7150	10400	9390	9020
angry	5820	6160	8520	10400	0,07	10500	10100	6790	7960	5600
sadness	11300	8580	8360	12200	10500	0,13	12900	8930	9210	10900
surprise	7680	7340	6110	7150	10100	12900	0,07	9360	8590	7130
grief	5030	7200	6550	10400	6790	8930	9360	0,1	9560	5270
disbelief	9730	4910	6480	9390	7960	9210	8950	9560	0,08	9120
satisfaction	2912	6060	6710	9020	5600	10900	7130	5270	9120	0,06

Euclidean distance	0-399	400-799	800-1220

Green cells of table 10 show the distances smaller than 400 pixels. Looking through the table we see that the distances between the same emotions in this table are very small values near to zero but these values are not exactly zero. These small values are the errors caused during recognition of the points. As we mentioned before because of some lighting problems in the recordings, some of the green points on the face are not recognized in a completely round shape and they may have different shapes in each frame. As a consequence the centers of the points which indicate the final locations of the points were not on the same place in each frame. That is why the distances between the same recordings are not zero.

Among all the distances between two different emotions the distance between 'satisfaction' and 'happiness' has the lowest value in the table (after distances between the same emotions) which is marked with green in our table. These two emotions are from the same family and they share lots of common facial features. The cells of the table with values between 800-1220 are marked with red and they show the big distances between emotions with very different facial features. The distance between 'happiness' and 'sadness' is one of the largest values of table 10. These two emotions belong to two different families with very different facial features. Another big distance of table 10 (marked with red) is the distances between 'satisfaction' and 'sadness'. These two emotions have also very different facial features. Other cells of our table are marked with blue which indicate the distance between 400 and 800. As you see using these distances we can define some clusters but for more features and clusters we need more information.

To understand the distribution of the distances and to see if we can make any conclusions based on the Euclidean Distances we calculate the mean and the standard deviations of the distances between emotions. The mean of a data set is simply the arithmetic average of the values in the set by summing the values and dividing by the number of values.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The standard deviation of a data set is the square root of the arithmetic average of the squared differences between the values and the mean.

$$s = \left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{\frac{1}{2}}$$

We calculate these values for distances between emotions and they are given in table 11. As it is shown in the table 11 the mean value of the distances between emotion 'sadness' and other emotions is bigger than the other mean values. Distance between 'sadness' and other emotions in our data forms an outline. Looking through the table 11 we notice that Euclidean distances of emotion 'sadness' with other emotions are also big. It confirms that the average value of distances between 'sadness' and other emotions is higher than the average values of distances between other emotions.

Table 11: Distribution of distances per emotion .

Emotion	Mean	Standard deviation
Happiness	7351	2598
Ironic smile	6456	1185
Astonishment	7004	1201
Understanding	9165	1808
Angry	7983	2001
Sadness	10320	1641
Surprise	8484	2064
Grief	7676	1950
Disbelief	8367	1644
satisfaction	6969	2408

If we compare the distances between emotions from table 10 with the results from Whissell database we got for a large part almost the same results. In Whissell dictionary thousands of words are rated in terms of two important factors: the word's pleasantness and their arousal levels (Gregory). The levels are from 1 to 3. We found these two levels for our labels and they are shown in table 12:

Table 12: The emotion's pleasantness level and arousal level.

Emotion	Pleasantness level	Arousal level
Happiness	3.0000	2.8000
Ironic smile	2.1111	1.5455
Understand	2.4286	1.8333
Angry	1.0000	2.5000
Sadness	1.3750	1.4286
Surprise	2.8571	2.6000
Grief	1.2500	2.0000
Satisfaction	2.5556	1.2500

The labels according to the levels of Whissell database are plotted in the figure 33. We calculated the Euclidean distance for the labels from Whissell database to compare them with the distances of table 10. Some of our labels ('astonishment' and 'disbelief') were not found in the Whissell dictionary and that is why these two columns are colored grey in table 13. In this table we also marked the values with 3 colors in 3 categories. Green cells show small distances between emotions and red cells indicate the big distances between emotions with very different values of activation and valence.

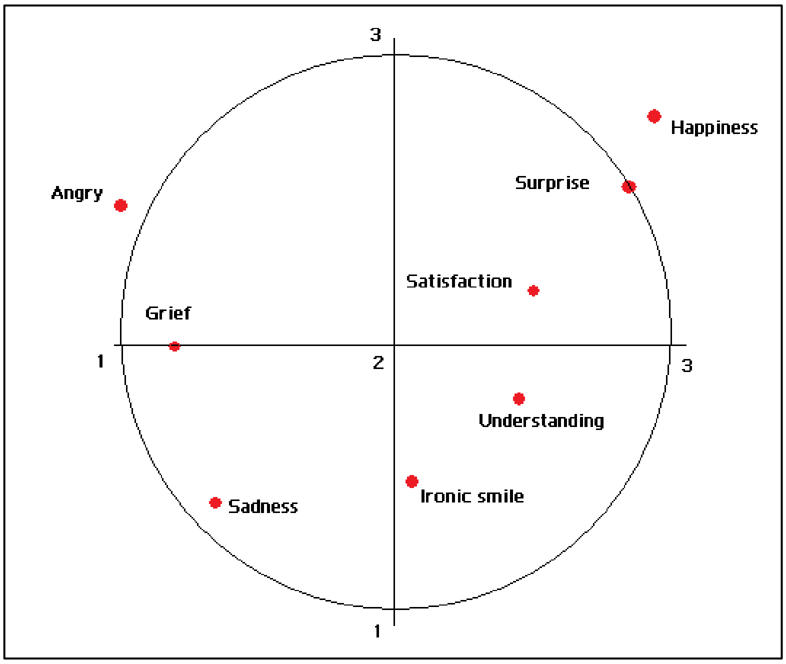
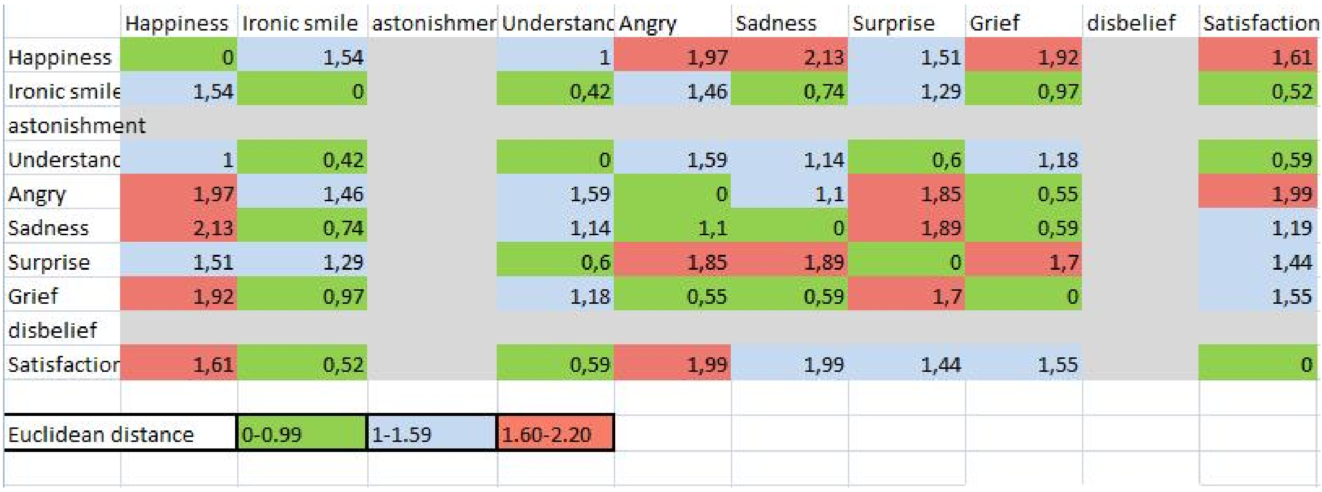


Figure 33: Labels from our table, rated using Whissell's database.

If we compare both tables together, we see some shared results. The value of distances between 'happiness' and 'sadness' from both tables are big and this according to our database and also Whissell dictionary means that these two emotions are very different from each other, based on facial features and the activation and the valence values. Also distance between 'surprise' and 'angry' in both tables belong to category red, which indicates the big distance between these emotions regarding to facial features and activation and valence levels. Not all the values from both tables show the same results. The lowest value of table 12 belongs to the distance between 'ironic smile' and 'understanding' and is marked with blue. Although not all the results from two tables are comparable but we can still see some shared results in two tables and using these results and extracted features from tracking points enable us to define some emotion clusters and use these cluster to label the recordings.

Table 13: Euclidean distances between different emotions from Whissell dictionary.



	Happiness	Ironic smile	astonishmer	Understanc	Angry	Sadness	Surprise	Grief	disbelief	Satisfaction
Happiness	0	1,54		1	1,97	2,13	1,51	1,92		1,61
Ironic smile	1,54	0		0,42	1,46	0,74	1,29	0,97		0,52
astonishment										
Understanc	1	0,42		0	1,59	1,14	0,6	1,18		0,59
Angry	1,97	1,46		1,59	0	1,1	1,85	0,55		1,99
Sadness	2,13	0,74		1,14	1,1	0	1,89	0,59		1,19
Surprise	1,51	1,29		0,6	1,85	1,89	0	1,7		1,44
Grief	1,92	0,97		1,18	0,55	0,59	1,7	0		1,55
disbelief										
Satisfactior	1,61	0,52		0,59	1,99	1,99	1,44	1,55		0

Euclidean distance	0-0.99	1-1.59	1.60-2.20

For defining enough features to recognize the emotions we need a much bigger database. During this project using our data, we extract a few features for some emotions and we are able to recognize some basic emotions, like recognizing ‘sadness’ from ‘happiness’ with 70% correctness and ‘anger’ from ‘surprise’ with 85%. For making a big rule base and recognizing more emotions we need more annotated data. We believe with more (annotated) data we would easily reach higher value for distinguishing all the 12 emotions from each other. This gives a basis for developing an automatic facial feature extraction. Using this method with a bigger data corpus enables us to extract enough facial features per emotion and make a fully automatic facial feature extraction system from face-to-face communications.

Chapter 7

7. Models of human face-to-face communication

In Chapter 1 we stressed the fact that emotions play an important role in human life. Humans are able express their emotional state by (non-) verbal communication. Facial expressions, gestures, and other body languages reveal the emotional state. In this thesis we focus on facial expressions and verbal content of speech. We defined important classes of emotional states. In this chapter we research the triggers of emotional states. We developed some probabilistic models and present them in the next sections.

7.1.Basic face-to-face communication models

In behaviorism as developed by Skinner (Skinner B. F.) human behavior is modeled by S-R (stimulus-response) paradigm. All behavior is triggered by stimuli. Those stimuli can be generated from the environment or internal. A movie of a dying young woman will definitely generate emotions of disgust, fear, compassion or aggression on the faces of observers. A funny thought will generate a smiling facial expression. The association S-R is realized by learning. The association stimuli-facial expression is mostly learned by imitation. On a very early age babies are stimulated to smiling parents. A smile of baby will be rewarded by enthusiasm of the parents. In this way a baby learns that positive reactions of parent can be evoked by a smile.

Figure 34: S-R mechanism between mother and baby.

In chapter 4 we analyzed our video recordings. We researched the relation between spoken words and facial expressions of the listener. We assume that the listener expresses his emotions by facial expressions. We are aware of the fact that some people have learnt not to reveal their real emotions. We also assume (as we explain in section 4.4) that those emotions are generated by the spoken text, by special key words, expressions of prosody (punctuation marks). The S-R association is context dependent. The generated expressions and its intensity depend of the situation, the context of the story and the current emotional state of the user (see figure 35).

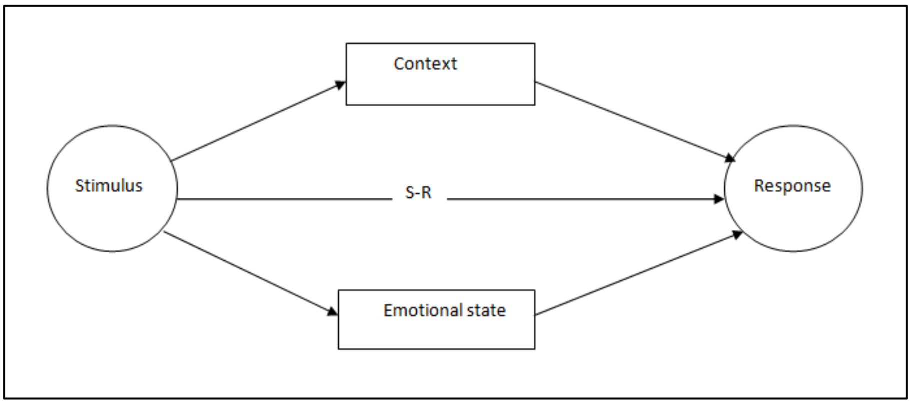


Figure 35: Context sensitive S-R model.

In table 14 we present the relation between keywords and facial expressions. We found different facial expressions. This can only be understood from the context. A complicating factor is that the context is not always represented by the keywords. Sometimes deep semantic processing is needed to understand the context. In the same way it is not always possible to explain the emotional state of the user by the last facial expression. A text is split up in emotional topics and based on the emotional topic of each part a facial expression is generated.

There are some more researches done for finding a relation between text triggers and facial expressions. One of them is the study done by E. Georgiana and D. Decheva (Georgiana E. and Decheva D., 2004). During their study they came to the conclusion that facial expressions are dependent not only on the words but also on the context of the conversation. So one word can mean different things according to the context of the dialogue. They compared the expressions that had the same triggers. Facial expressions are not identical because everybody reacts in different way in similar situation. In their results triggers are the same but the expressions are different.

Table 14: Example of facial expressions triggered by the keywords.

	Key word	**Facial expression**	**Key word**
Astonishment	*So, Immediately, fascinated .*	Anger	*Sadist, Crying, Annoying.*
Surprise	*Oh, No?, what happened?, Birthday party, Unusual .*	Satisfaction	*Oh, Happy, Nice,Yes, My darling, Happy, Delicious, Succeed.*
Sadness	*Dying, Sick, Troubled.*	Happiness	*Oh, Succeed, joyfully, Cooking book, Good night.*
Disbelief	*Strange, refuse, non sense, Awful, Doubtfully, Strange .*	Understanding	*Aha, OK, No problem.*
Disgust	*Cutting people.*	Ironic smile	*Ridiculously.*
Grief	*Banging, chicken pox.*	Regret	*Oh, No.*

We found that usually a facial expression is generated from a neutral facial expression. But in case of blended emotions and facial expressions we noticed that the users are able to generate successive facial expressions or fuse them together. A common example is an expression of (displaced) fear or fright which is usually followed by happiness and laughing.

In figure 36 we display all emotions and transitions between one of the emotions and the rest of them. This can be considered as a Markov model, with transition probabilities between emotional states. In figure 36 the transition probabilities between different emotional states are given. One of disadvantages of the Markov's model is that it displays only the states and transition between them, but not the triggers.

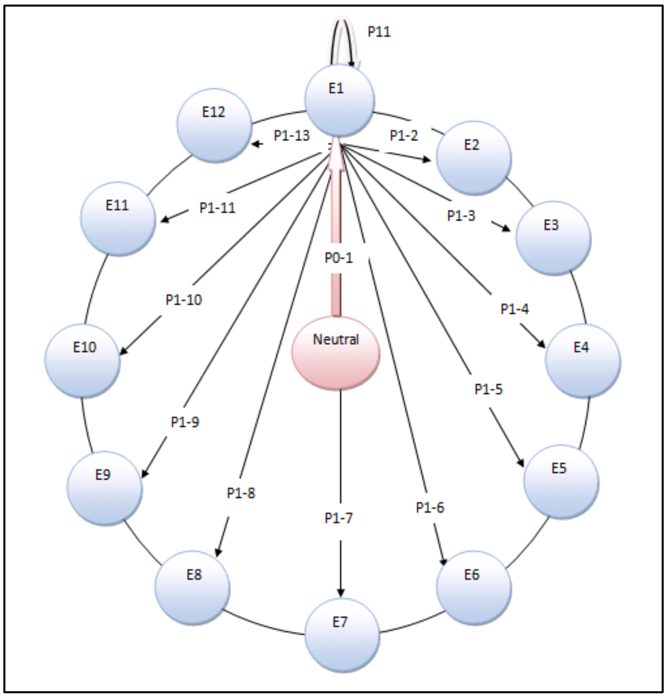


Figure 36: Markov model of emotional states.

7.2.Causal model of interactions

Most 2-dimensional models of emotions are based on Valence-Arousal model. Every emotional word can be measured on the valence/arousal scale as realized in the Whissell database (Gregory). These values can be considered as x-y coordinates. So all emotional words can be plotted as points in a two dimensional space. In figure 37 we plotted some of the emotional words. In the same way we can plot the 12 emotional states as points in the plane (see figure 38).

Now we define our first causal model. The vectors from figure 37 can be considered as displacement in figure 38. Let us assume that some emotional keywords are present in a text fragment (see figure 39). We represent these keywords as vectors in figure 40 and compute the average sum. Let us assume that the listener was in a neutral state. By listening to the text he will be transferred to an emotional state. We shift the neutral state by average sum vector and the closest emotional keyword is the assumed new emotional state as in is shown in figure 41. We assume that the emotional states are not single points but clusters with the labeling point as bary center. We assume a Gaussian distribution around the centre expressing the location of the specific emotional state.

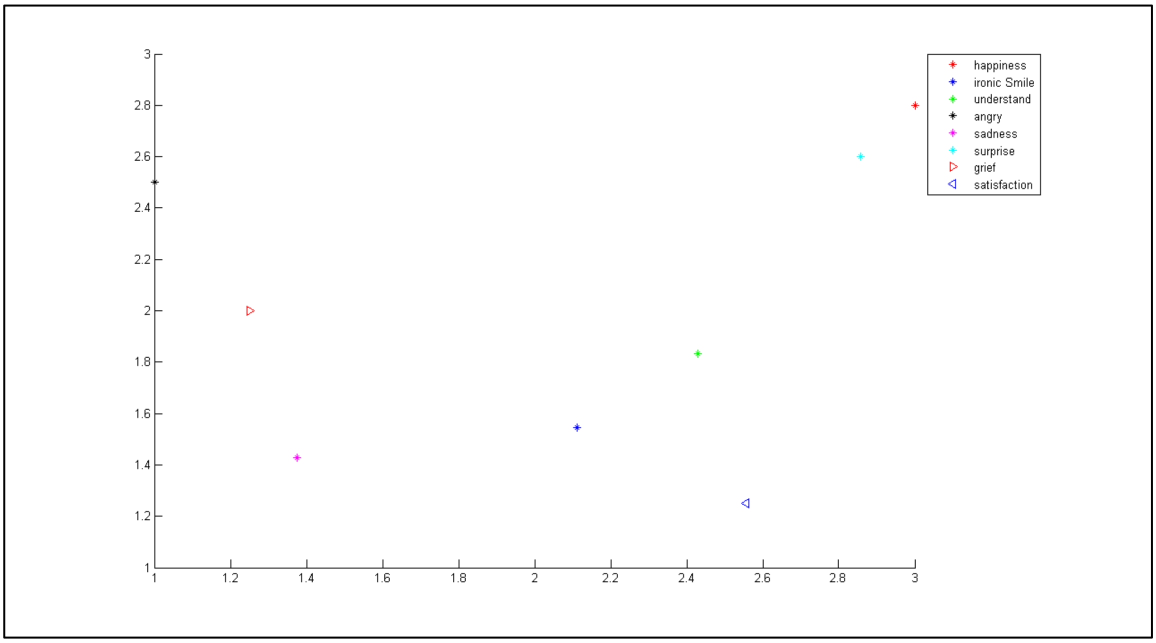


Figure 37: Emotional states.

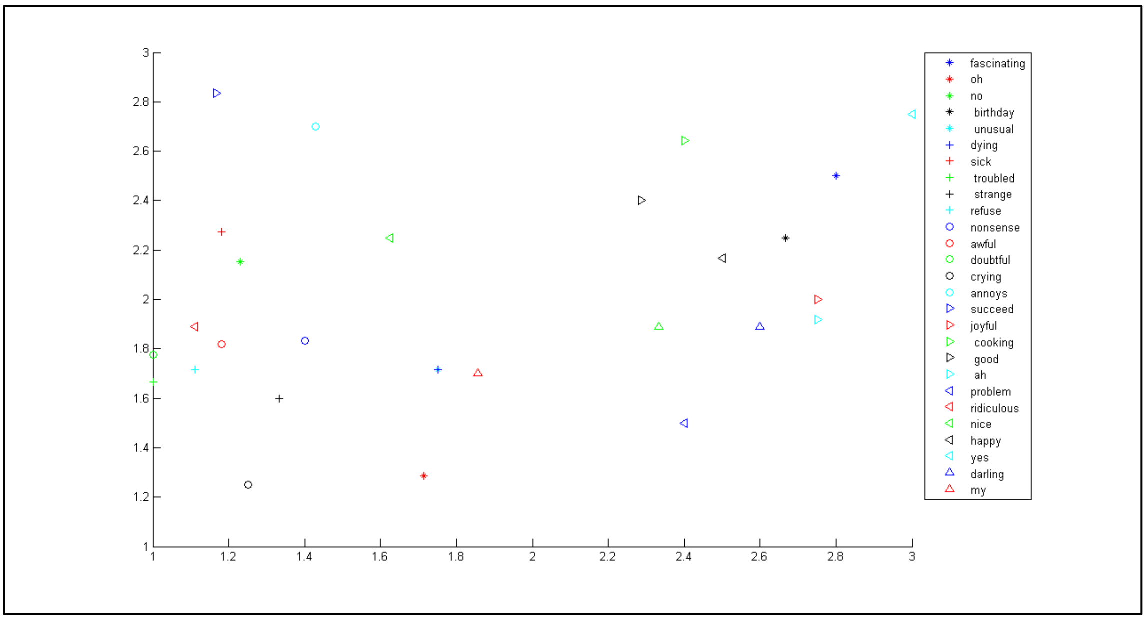


Figure 39: Emotional words.

— Now, I would like to make a fancy-cake for my father, for his birthday-party.

— Fancy cake? Oh my God, Gabrysia, I would suggest you don't do it.

Figure 38: A text fragment with some emotional words.

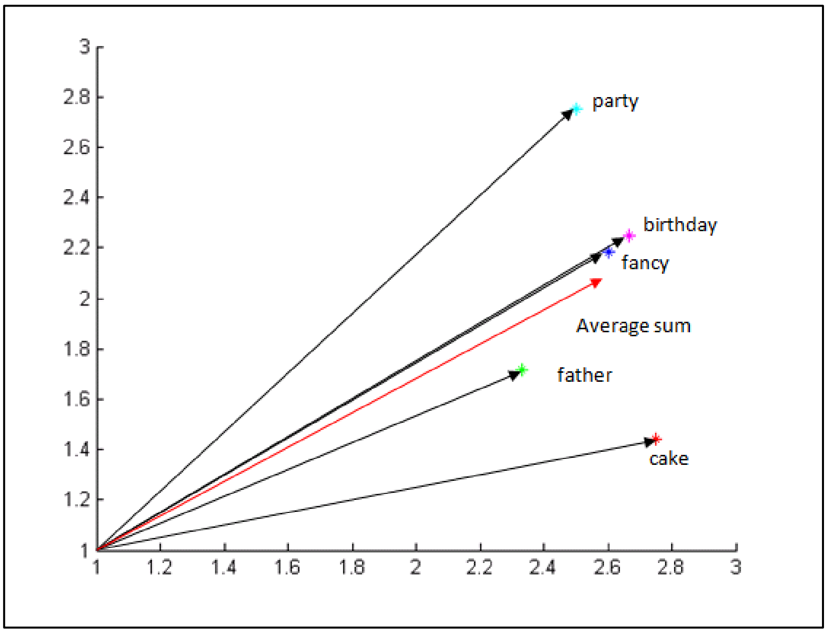


Figure 40: Emotional words of one sentence and the average sum of them presented as vectors.

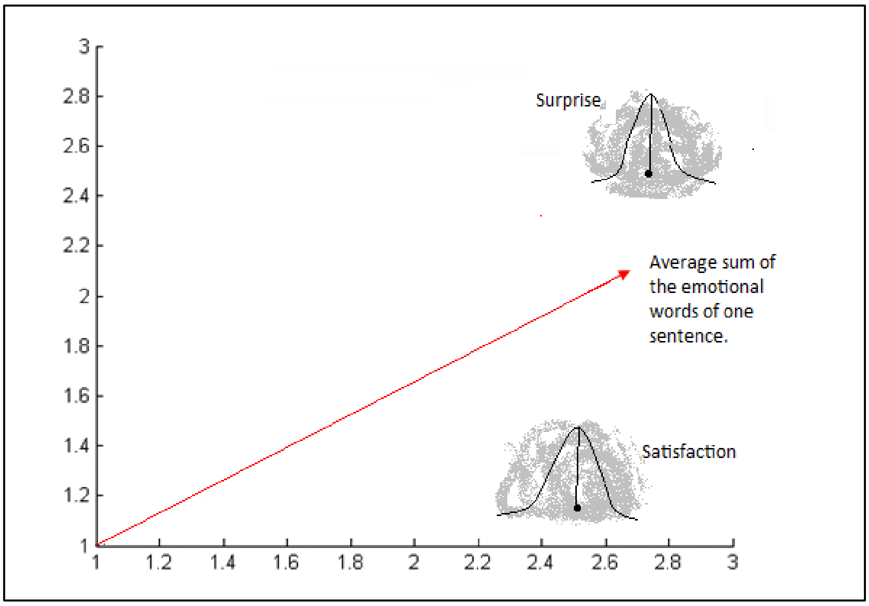


Figure 41: Average sum of the emotional words and the Gaussian distribution of emotional states 'surprise' and ' satisfaction'.

7.3.Bayesian model of interaction

Finally we can model our S-R reaction as a Bayesian network. We consider the model as displayed in figure 35. The emotional states are measured on a nominal scale. But we can use the valence and arousal score of every sate to get a metric score. In the same way the context can be measured on the valence and arousal score. The valence and arousal scale can be split up in 5-7 classes. Then we can define *conditional probability tables* between the triggers and response. The probabilities in the CPT tables can be defined by experts or computed. Unfortunately our database is very limited. So the entries are defined on opinions of experts. In fact the relation between S-R is modeled by a (linear) function. It is assumed that increasing valence or arousal generates emotional states with increasing valence and arousal. In this way we define the entries in the diagonal of the TCP tables. Next we add some random noise to fill the other entries to model the remanding state transition.

7.4.Validation of developed interaction models

In this section we will give an empirical rounding of our developed models. In chapter 4 we analyzed our annotated video recordings. Rates were requested to analyze the annotated videos. Unfortunately the level of agreement between annotators was low. This is not exceptional but also found by other researchers (Reidsma D., 2008) (Steininger S., 2006). One of the main reasons is that interpretation of emotion is very subjective. The same video can generate different emotional reactions. Every annotator has his own view on the world, his own history, education and experiences. As a result the interpretation and corresponding annotation can be very subjective. In case of famous Clinton trials 50% of the people found that Clinton was lying and the other half found he was speaking the truth. Our daily interaction suffers from the same ambiguities. But usually during an interaction speaker and listener solve their ambiguities by asking questions or testing hypothesis.

Anna Wojdel analyzed the recordings and found the relation between triggers and facial expression (see chapter 4). Analyzing our annotated recording we get also the same results. It proves that in about 55% of the cases the emotional expression could be explained by the emotional keywords in the corresponding fragment of speech. So using our S-R model or geometrical model in about half of the cases we are able to predict the right facial expression. In other half of the cases either no emotional keywords are available, or the emotion is triggered by the context. At this moment we analyzed only the text, no other nonverbal features from the context are taken into account. By analyzing some cases it proves that a lot of common sense knowledge is used to understand the generated facial expression. We refer to the thesis of I. van Willegen where he reports that modeling common sense knowledge is one of the challenging open research problems in Artificial Intelligence and far beyond scope of this thesis.

Chapter 8

8. Conclusions and future work

Figure 42: At 1001th night sultan decides to spare the life of Shahrdzad (Lang, 1898).

In chapter one, we talked about Shahrdzad and the bitter sultan. At the 1001th night Shahrdzad's story finally reached its end. The sultan had not only been entertained but wisely educated in morality and kindness by Shahrdzad. He decided to spare the life of Shahrdzad who became his Queen. I loved this part of the story. The moment that sultan finally realize that after 1001 nights and 1001 stories they have shared so many emotions that he cannot even remember why he was so bitter and why he wanted to execute Shahrdzad. All he could remember at that moment was the nice stories of her. The human emotions may not disappear from our life. We have to and we would find a way to adapt the new communication system to our self, our emotions and our needs.

As is mentioned in other chapters of this report there are lots of researches going on at this field. The main goal of our research was to extract facial expressions from video recordings in face-to-face communication. In chapter 1 we formulated a number of research questions to answer during this thesis project. In this chapter we give answers to those questions.

- *Can we successfully create a protocol for the creation of a multimodal database?*

  As it is explained in chapter 3 we have successfully created a recording protocol for a multimodal database (The New Delft University of Technology Audio-Visual Speech) and we have successfully setup a new multimodal database. This database and its content should be made available to as many researchers as possible to compare the achieved results by them.

- *Can we compare the different presumptions of annotators of a dataset and calculate a level of agreement between them?*

  After the annotators annotated the recordings with our offered labels we analyzed the corpus. We compared the annotated parts with each other and calculated the level of agreements between them. After some more inspection we found out that some of our offered labels are from same family and it is very confusing for the annotators. We categorized our labels and put the labels like 'astonishment' and 'surprise' in the same category and calculate the level of agreement for the second time. As we had expected this time the level of agreement between the annotators was higher than the first time and in some recordings they reached the highest possible level of agreement which is equal to 1. But unfortunately this high level was reached not very often and generally the level of agreement between the annotators was low as it was found by other researchers (Reidsma D., 2008) (Steidl S., 2008). The low levels are caused among others by different view and experiences of each annotator. Each video can generate different emotional reactions. .

- *Can we track the points on the contour of the lip and eyebrows to extract facial features per each emotion?*

  As it is explained in chapter 6 after visual inspection of annotated data we select the important points namely points on the contour of mouth and eyebrows. Following these points we extract some facial features and define a feature set per each emotion. For making a bigger feature sets we need more carefully annotated recordings to extract more features per each emotion.

- *Can we define some emotion clusters using the extracted features?*

  After following the movements of the points and extracting the features we noticed that there are some emotions clusters formed. Following the points during a recording based on the movements we ended up in different areas. Each area is an emotion cluster and we use these clusters to label the emotions. The extracted features guide us to one of these emotion clusters and enable us to recognize the emotion and label it. If the extracted features guide us to an area between the clusters calculating the distance to each cluster give us the probabilities for each cluster.

- *Can we build an automatic emotion extractor for the face-to-face communication?*
  Having more data enables us to extract more features and define emotion clusters for building a full automatic facial emotion extractor from the face-to-face communications. At this moment we have not enough data to define features and emotion clusters for all the labeled emotions and train our model.

8.1. Concluding remarks

The final goal of the MMI project is to assess facial expressions in an automated way. This thesis is an intermediate step. The extraction process is facilitated by the green marker on the face. Human observes are used to localize and label facial expressions and get insight how many facial expressions are used, what triggers a facial expression in our daily life communication. The results of this thesis produce necessary input for the next steps, the automated extraction of facial expressions from video recording of human interacting.

During this project we have focused on construction of a model for extracting emotions from face-to-face communication. We gathered theoretical information about the human emotions which is described in chapter 1. Several researchers tried to extract and recognize human emotions from (sequence of) images. The used corpus for this project was of great importance. For extracting the correct features we have to find the proper data to work with. As it is explained in chapter 3 after some research when we couldn't find the proper data we decided to build our own database. For this purpose we made a protocol for recording the face-to-face communication and start the recordings based on our scenarios. Next we arranged an experiment with 3 nonprofessional annotators to annotate our recordings. Annotating the recording is a very important and time consuming task and not carefully annotated recordings can strongly influence the results. After annotating the recordings to compare the annotation of each annotator with each other we defined a level of agreement to understand how much people differ in presuming an emotion. Calculating the level of agreement gives us information about the difficulty of annotation task, vagueness of the shown emotions in the recordings and amount of mistakes annotators made. In case of non-spontaneous labeling it also shows the suitability of the offered labels to annotators. In some recordings we reached the highest value for the level of agreement between the annotators. But because of subjectivities of each annotator the reached level of agreement was most of the time not very high.

After acquiring the annotated data and gathering knowledge about the subject we studied the methods used by other researchers for extracting facial features. We developed a model to extract the facial features by following the movements of different elements on the human's face and for simplifying the tracking we used 31 colored markers on the face. Based on the extracted features and Euclidean distances between different emotions we found some emotion clusters. These clusters are used for assigning a label to new recordings. The extracted features guide us to the one of these clusters and enable us to recognize the emotion and assign a label to it. This gives a basis for developing an automatic facial features extraction tool. A big and carefully annotated corpus helps us to extract enough facial features per emotion and make a fully automatic facial feature extraction tool from face-to-face communications.

8.2. Future work

As mentioned before we need much more annotated data to extract more facial features from them to define the emotion clusters and finally build an automated emotion extraction tool for the face-to-face communication. At this moment we have very little features based on the following some points on the face and comparing the Euclidean distance between them. To follow more points on face for getting more facial features we need new recordings. Making new recordings is not difficult and doesn’t take long as we made a protocol for it, but annotating the recording is very time consuming. Annotating the data should be done very carefully and according to the annotation protocol. In future adding more recordings and also adding more labels to the label list is possible. Also labeling spontaneous can replace the labeling according to a list. But we should be very careful with it, because it may cause lots of confusing between the annotators. Having more annotators to annotate the data also improves the results.

During our project we filled our database with frontal view and profile view of the actors. But as we used the recording made by Anna and her data just contains frontal view we couldn’t extract any features from profile view. There is lot of information on the profile view of the face and some of the changes are even more visible on the profile view than on frontal view. Using both views gives us more extracted features and help us to define clusters and recognize the emotions better.

Using speech recognition methods would also provide us more features. The vocal aspect of communication contains lots of information. Adding a speech recognition part to this tool helps us to extract more features and a bigger set of extracted features per emotion helps to define the emotion cluster more accurate and improve the results.

The points extraction based on the colored marker during this project was just to simplify the tracking and in the near future it can be replaced by other methods.

Bibliographies

Anvil *http://www.anvil-software.de/*.

Bargeron D. Gupta A., Gurdin J. and Sanocki E. *Annotation for streaming Video on the web: System Design and Usage Studies.* - 1999.

Beckman M.E. and Elam G.A., *Guidelines for ToBI labelling.* - 1997.

Benesh *http://www.benesh.org/BNHome.html* - Benesh Institute.

Black A. W., Taylor, P., and Caley, R. *The Festival Speech Synthesis System, System documentation.* Centre for Speech Technology Research University of Edinburgh - 2002.

Cassell J. Vilhjálmsson H. H. , Bickmore T. *BEAT: the Behavior Expression Animation Toolkit.* - 2000.

Chitu A. Vulpen van M., Takapoui P. and Rothkrantz L.J.M. *Building a Dutch Multimodal Corpus for Emotion Recognition Building a Dutch Multimodal Corpus for Emotion Recognition* - 2008.

DeCarlo D. Stone M. *The Rutgers University Talking Head: RUTH* - 2003.

DeCarlo D. Stone M., Corey Revilla and Jennifer J. Venditti, *Specifying and animating facial signals for discourse in embodied conversational agents* - 2004.

Desmet P. M. A. *Designing Emotions* - Delft University of Technology, Delft , 2002.

Ekman P, *Emotion in the human face* - 1982.

Ekman P. *An argument for basic emotions, cognition & emotion* 6,196-75,192-20 - 1992.

Ekman P. and Friesen, W. V. *Unmasking the Face* - Prentice Hall. - 1975.

Ekman P. *Facial expression of emotion* - 1992.

Ekman P. Lepenies K., and Ploog, D. *About brows: emotional and conversational signals. Human ethology:* claims and limits of a new disipline: contributions to the Colloquium, - 1979. - pp. pages 169–248.

Ekman P. Oster H. *Universals and cultural differences in facial expressions of emotion.* Nebr. Symp. Motiv - 1972. - pp.207-83 .

ELAN *http://www.let.ru.nl/sign-lang/ECHO/ELAN/ELAN_intro.html*.

Friesen V. and Ekman P. *The facial action coding system: A technique for measurement of Facial Movement. Consulting Psychologists Press*, Palo Alto - 1978.

Georgiana E. and Decheva D., *Talking face,* TU Delft - June 2004.

Gower J. C. *Euclidean Distance Matrix*.

Gregory N. Purdue U. *Assessing the Affective Aspect of Languaging: The Development of Software for Public Relations* - Hammond, USA .

Gussenhoven C. *Transcription of Dutch Intonation.* In Jun, Sun-Ah (ed.), Prosodic Typology: The Phonology of Intonation and Phrasing.Oxford: Oxford University Press. - 2005. - pp. 118-145.

Ismail M., *Video content analysis & aggression detection system for a train environment* - Delft - 2007.

Kappa *http://cosmion.net/jeroen/software/kappa/*.

Klein M., *MATE DELIVERABLE 1.1. Supported Coding Schemes* - July 1998.

Krippendorff K. *Content Analysis: An Introduction to Its Methodology* - Beverly Hills, CA : Sage Publications - 1980.

Lang A. *The Arabian Nights* - Biblio Bazaar, LLC - 1898.

Marsi E. Ferdi van Rooden, *Expressing Uncertainty with a Talking Head in a Multimodal Question-Answering System* - 2007.

Mehrabian Albert, and Ferris, Susan R. *Inference of Attitudes from Nonverbal Communication in Two Channels*. Journal of Consulting Psychology, June 1967. - Vol. 31, No. 3,pp. 248-258.

Pantic M. Rothkrantz L.J.M. E*xpert system for automatic analysis of facial expressions* - Faculty of Information Technology and Systems, Department of Knowledge Based Systems, Delft University of Technology,P.O. Box 356, 2600 AJ Delft, The Netherlands, October 2000.

Pantic M. Rothkrantz L.J.M. *Toward an Affect-Sensitive Multimodal Human–Computer Interaction* - Proceedings of the IEEE, vol. 91, no. 9, pp. 1370-1390, September 2003.

Pantic M. Student Member, IEEE, and Rothkrantz L.J.M. *Automatic Analysis of Facial Expressions:The State of the Art*.

Reidsma D. *Annotations and subjective machines of annotators, embodied agents, users and other humans* - 2008.

Shlens J. *Tutorial on Principal Component Analysis* - Institute for Nonlinear Science, University of California, San Diego - 2005.

Silverman K. Beckman M. *TOBI: A standard for* - 1992.

Skinner B. F. *The Behavior of Organisms: An Experimental Analysis.* New York: Appleton-Century-Crofts,. - Vol. 1938.

Smartkom, *http://smartkom.dfki.de/start_en.html*.

Steidl S. *Automatic Classification of Emotion-related User States in Spontaneous Children's Speech* - Universität Erlangen-Nürnberg - 2008.

Steininger S. Schiel F., Rabold S. *Annotation of multimodal data* - 2006.

Steininger S. *Translation of language and labeling and labeling of emotion and gesture in smartkom* - 2001.

Stone M. DeCarlo D. *Crafting the Illusion of Meaning: Template-based Specification of Embodied Conversational Behavior* - 2003.

ToBi *http://www.ling.ohio-state.edu/~tobi/*.

Wojdel A. *Knowledge driven facial modeling* - Delft - 2005.

Appendices

A: Annotation protocol

You are about to annotate some recordings using the annotation tool ELAN. To start with annotating you watch some recordings and when you notice an expression, select the frame with the expression and define the beginning and the end of the expression. Then assign a label to selected frames from the list we offer to you. On the list there are 12 different labels and each label is presenting a facial expression. (The next 4 steps are already done. You can start with step 5.)

1. To begin with annotation first open the new document you want to annotate in ELAN main menu using *File->New*.

2. Create a new Linguistic Type by Choosing *Type->Add New Linguistic Type*. Enter a name for the new type and click the *Add* button.

3. Now you should define a new Tier using *Tier->Add New Tier*. Enter a name for the new tier and click *Add* button.

4. The new tier appears in the low left side in the ELAN window. Double click it to make it active.

5. Use the media player control button to start and stop the media player. Using the mouse you can select a segment of the video on the timeline viewer.

6. Create the annotation by double clicking on the selected segment. Enter the edit box that appears and give a proper label then confirm it by hitting *Ctrl + Enter* .

Congratulation! Now you have annotated the selected segment.

If you are not satisfied with the start point and end point of the annotation it can be modified by selecting the right section and hitting *Ctrl + Enter* or right clicking and choosing *Modify Annotation Time*. The value of the annotation can be edited by double clicking the annotation. Sometimes you may see more than one expression simultaneously. In that case start with the expression you see stronger and then assign the second label also.

You can repeat these steps as long as you want till you are satisfied with it.
Good luck!

At the back side there is list with all the labels and feature of them. Also a picture of each emotion is also assigned.

Here is the list with label on it.

1. Astonishment raised eye-brows and eyes wide open
2. Surprise raised eye-brows
3. Sadness lowered corner of the mouth, raised chin
4. Disbelief lower eye-brows and mouth slightly stretched
5. Regret tightened and stretched mouth
6. Grief raised inner eye-brows
7. Anger lowered eye-brows
8. Disgust wrinkled nose
9. Happiness open mouth, raised corner of mouth and raised cheeks
10. Understanding withdrawn and lifted up head, mouth open and slightly raised eye-brows
11. Satisfaction slightly raised chin and corners of the mouth
12. Ironic smile raised upper lip and corner of the mouth

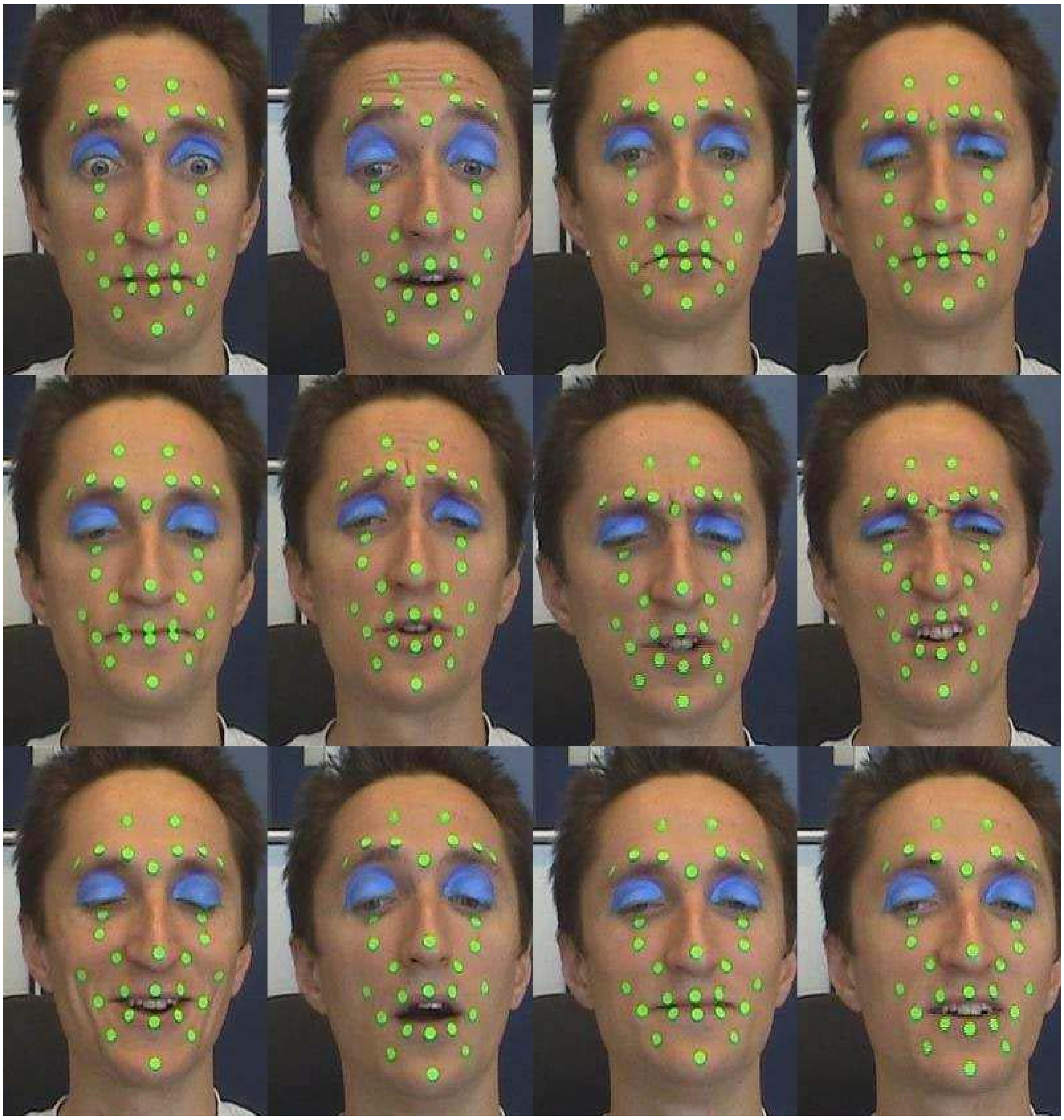


Figure 43: Template expressions.

B: The annotated recordings

Annotator1

Anna1

Labels	TC	Triggers
Disbelief	00:00:11.320 - 00:00:13.080	My name is ...
Disbelief	00:00:19.130 - 00:00:20.940	No, that ..
Surprise	00:00:26.810 - 00:00:28.080	Oh, that
Satisfaction	00:00:35.330 - 00:00:37.890	Oh, I understand
Happiness	00:00:48.740 - 00:00:51.720	oh so, have you also noticed it
Disbelief	00:01:04.450 - 00:01:05.300	strange
Understanding	00:01:17.010 - 00:01:18.590	mysterious
Surprise	00:01:44.030 - 00:01:45.050	No?
Understanding	00:01:52.520 - 00:01:53.440	Why aren't you
Satisfaction	00:01:53.480 - 00:01:55.000	Father was speechless
Disbelief	00:02:12.030 - 00:02:14.020	if you don't believe me, please..
Satisfaction	00:02:32.010 - 00:02:34.990	We will do our best. We will try to consider your wishes.

Anna2

Labels	**TC**	**Triggers**
Disbelief	00:00:10.990 - 00:00:12.010	Father was obstinately refusing
Disbelief	00:00:15.000 - 00:00:16.010	They told you everything
Surprise	00:00:21.990 - 00:00:22.990	face expression
Disbelief	00:00:25.420 - 00:00:26.660	He had brows drown together
Disbelief	00:00:39.420 - 00:00:40.600	I have to be here
Satisfaction	00:00:41.520 - 00:00:44.010	what those doctors may find out
Anger	00:01:05.010 - 00:01:06.560	Most of the surgeons are sadists
Surprise	00:01:19.000 - 00:01:20.010	He came immediately
Understanding	00:01:26.560 - 00:01:28.010	the desperate father can do
Understanding	00:01:43.000 - 00:01:44.696	That was a truth

Anna3

Labels	**TC**	**Triggers**
Surprise	00:00:05.550 - 00:00:06.720	You called
Surprise	00:00:27.010 - 00:00:28.000	She drank seed flax
Grief	00:00:40.550 - 00:00:41.580	she should not get excited
Disbelief	00:00:49.460 - 00:00:50.550	about three weeks
Surprise	00:00:57.010 - 00:00:58.000	With everything
Satisfaction	00:01:02.010 - 00:01:03.540	you don't see it

Anna4

Labels	TC	Triggers
Satisfaction	00:00:07.500 - 00:00:09.610	she looked happy
Anger	00:00:24.010 - 00:00:25.010	she was crying
Satisfaction	00:00:43.010 - 00:00:44.010	OK?
Satisfaction	00:00:55.450 - 00:00:56.580	frightening me
Disbelief	00:00:56.630 - 00:00:58.010	nonsense about chicken
Ironic smile	00:01:10.000 - 00:01:12.010	Darling, you can have three chickens
Happiness	00:01:12.080 - 00:01:15.510	Sleep well, I will be back late
Regret	00:01:49.010 - 00:01:50.580	Chicken-pox
Satisfaction	00:02:14.030 - 00:02:15.030	Gabrysia began to bite on her fingers

Anna5

Labels	TC	Triggers
Satisfaction	00:00:11.000 - 00:00:12.010	she groaned when she saw
Astonishment	00:00:21.010 - 00:00:23.510	what happened this night
Ironic smile	00:00:23.580 - 00:00:25.530	could not sleep
Disgust	00:00:33.010 - 00:00:34.010	I heard even you snoring
Ironic smile	00:00:37.470 - 00:00:39.000	you should not snore
Surprise	00:00:41.020 - 00:00:43.000	has forgotten
Disbelief	00:00:46.010 - 00:00:49.000	She is an awful woman
Astonishment	00:01:01.000 - 00:01:03.490	at her place tonight?
Surprise	00:01:15.010 - 00:01:17.620	was staring at Gabrysia
Anger	00:01:47.010 - 00:01:48.020	and annoying crack
Surprise	00:02:01.460 - 00:02:03.010	Watch your manners
Anger	00:02:06.010 - 00:02:08.010	she can hear everything

Anna6

Labels	TC	Triggers
Sadness	00:00:11.450 - 00:00:12.560	independent expression (no trigger)
Satisfaction	00:00:14.010 - 00:00:15.000	blushed ridiculously
Satisfaction	00:00:17.510 - 00:00:18.610	That's very nice
Satisfaction	00:00:29.000 - 00:00:30.020	Oh, yes
Astonishment	00:01:02.020 - 00:01:03.480	Really?
Satisfaction	00:01:09.440 - 00:01:11.470	Water!!!
Disbelief	00:01:13.010 - 00:01:14.500	I'm sick. I'm dying.
Astonishment	00:01:33.010 - 00:01:34.000	so?
Disbelief	00:01:43.020 - 00:01:44.440	to beat red-haired severally

Anna7

Labels	TC	Triggers
Disbelief	00:00:34.000 - 00:00:35.010	you did not hear me at all
Satisfaction	00:00:40.010 - 00:00:42.000	my darling
Understanding	00:00:59.000 - 00:01:00.000	father struggled with his urge
Satisfaction	00:01:06.430 - 00:01:07.520	I can cook
Satisfaction	00:01:26.010 - 00:01:28.000	A cooking-book

Disbelief	00:01:37.000 - 00:01:38.010	my darling
Understanding	00:01:43.010 - 00:01:44.000	he put cooking-book aside

Anna8

Labels	TC	Triggers
Anger	00:00:14.540 - 00:00:16.510	Men are mean animals
Satisfaction	00:00:19.480 - 00:00:21.010	that's right
Disbelief	00:00:24.990 - 00:00:26.510	also mean
Satisfaction	00:00:34.500 - 00:00:36.500	full of nuts and all other things
Understanding	00:00:40.010 - 00:00:41.000	asked doubtfully
Astonishment	00:02:00.000 - 00:02:00.990	fascinated by our father
Satisfaction	00:02:11.010 - 00:02:13.000	terrible afraid of mom
Surprise	00:02:31.000 - 00:02:31.990	rascal
Satisfaction	00:02:41.990 - 00:02:43.000	very surprised
Satisfaction	00:03:04.990 - 00:03:06.010	joyfully
Ironic smile	00:03:07.020 - 00:03:08.010	Aspazja from Millet

Anna9

Labels	TC	Triggers
Disbelief	00:00:03.990 - 00:00:05.510	this time
Ironic Smile	00:00:06.000 - 00:00:07.000	nothing unusual
Satisfaction	00:00:08.000 - 00:00:09.000	delicious
Surprise	00:00:11.000 - 00:00:12.000	birthday-party
Surprise	00:01:03.510 - 00:01:04.500	bubble
Satisfaction	00:02:11.990 - 00:02:13.000	succeed

Anna10

Labels	TC	Triggers
Satisfaction	00:00:12.480 - 00:00:13.560	Colors of earth
Satisfaction	00:00:16.990 - 00:00:18.000	grumbled
Understanding	00:00:47.000 - 00:00:48.000	in these clothes
Satisfaction	00:02:03.010 - 00:02:05.980	full of style
Happiness	00:02:12.000 - 00:02:13.010	look like Clark Gable
Happiness	00:02:28.010 - 00:02:30.000	so, don't say anything

Annotator 2

Anna 1

Labels	TC	Triggers
Surprise	00:00:02.880 - 00:00:06.250	We do not know each other
Surprise	00:00:18.050 - 00:00:19.010	This door
Understanding	00:00:26.830 - 00:00:28.090	Oh, that
Ironic smile	00:00:35.330 - 00:00:37.890	Oh, I understand
Understanding	00:00:40.840 - 00:00:42.700	it blows from downstairs...
Happiness	00:00:48.740 - 00:00:51.720	oh so, have you also noticed it
Disbelief	00:01:04.450 - 00:01:05.300	strange
Understanding	00:01:17.010 - 00:01:18.590	mysterious
Surprise	00:01:44.030 - 00:01:45.040	No?
Surprise	00:01:52.520 - 00:01:53.440	Why aren't you
Satisfaction	00:01:53.480 - 00:01:55.000	Father was speechless
Grief	00:02:09.010 - 00:02:10.000	banging doors
Satisfaction	00:02:32.010 - 00:02:34.990	We will do our best.

Anna2

Labels	TC	Triggers
Disbelief	00:00:10.990 - 00:00:12.010	Father was obstinately refusing
Disbelief	00:00:15.010 - 00:00:15.990	They told you everything
Disbelief	00:00:25.420 - 00:00:26.660	He had brows drown together
Grief	00:00:39.470 - 00:00:40.620	I have to be here
Satisfaction	00:00:41.520 - 00:00:44.010	what those doctors may find out
Disgust	00:01:05.010 - 00:01:06.560	He finds pleasure in cutting people
Surprise	00:01:19.000 - 00:01:20.010	He came immediately
Understanding	00:01:26.560 - 00:01:28.010	the desperate father can do

Anna3

Labels	TC	Triggers
Disbelief	00:00:05.550 - 00:00:06.720	You called
Astonishment	00:00:27.010 - 00:00:28.000	She drank seed flax
Surprise	00:00:39.010 - 00:00:40.530	cut out half of her stomach
Surprise	00:00:57.010 - 00:00:58.000	With everything
Ironic smile	00:01:02.010 - 00:01:03.540	you don't see it

Anna4

Labels	TC	Triggers
Satisfaction	00:00:07.500 - 00:00:09.610	she looked happy
Satisfaction	00:00:43.010 - 00:00:44.010	OK?
Anger	00:00:55.450 - 00:00:56.580	frightening me
Disbelief	00:00:56.630 - 00:00:58.010	nonsense about chicken
Ironic smile	00:01:10.000 - 00:01:12.010	My darling, you can have three chickens

Happiness	00:01:12.080 - 00:01:15.510	Sleep well, I will be back late
Anger	00:01:49.010 - 00:01:50.580	Chicken-pox
Understanding	00:02:14.030 - 00:02:15.030	Gabrysia began to bite on her fingers

Anna5

Labels	TC	Triggers
Satisfaction	00:00:11.000 - 00:00:12.010	she groaned when she saw
Surprise	00:00:21.010 - 00:00:23.510	what happened this night
Ironic smile	00:00:23.580 - 00:00:25.530	could not sleep
Disbelief	00:00:33.010 - 00:00:34.010	I heard even you snoring
Ironic smile	00:00:37.470 - 00:00:39.000	you should not snore
Surprise	00:00:41.020 - 00:00:43.000	has forgotten
Disbelief	00:00:46.010 - 00:00:49.000	She is an awful woman
Astonishment	00:01:01.000 - 00:01:03.490	at her place tonight?
Surprise	00:01:15.010 - 00:01:17.620	was staring at Gabrysia
Disbelief	00:01:47.010 - 00:01:48.020	and annoying crack
Astonishment	00:02:01.460 - 00:02:03.010	Watch your manners
Surprise	00:02:34.010 - 00:02:35.520	At midnight

Anna6

Labels	TC	Triggers
Sadness	00:00:11.450 – 00:00:12.560	(no trigger) independent expression
Ironic smile	00:00:14.010 – 00:00:15.000	blushed ridiculously
Ironic smile	00:00:17.510 – 00:00:18.610	That's very nice
Satisfaction	00:00:29.000 – 00:00:30.020	Oh, yes
Surprise	00:01:02.020 – 00:01:03.480	Really?
Satisfaction	00:01:09.440 – 00:01:11.470	a dramatic voice resounded
Disbelief	00:01:13.010 – 00:01:14.500	I'm sick. I'm dying.
Astonishment	00:01:33.010 – 00:01:34.000	so?
Disbelief	00:01:43.020 – 00:01:44.440	to beat red-haired

Anna7

Labels	TC	Triggers
Disbelief	00:00:19.000 – 00:00:20.010	What a pity
Disbelief	00:00:34.000 – 00:00:35.010	you did not hear me at all
Ironic smile	00:00:40.010 – 00:00:42.000	my darling
Satisfaction	00:00:59.000 – 00:01:00.000	father struggled with his urge
Satisfaction	00:01:06.430 – 00:01:07.520	I can cook
Satisfaction	00:01:26.010 – 00:01:28.000	A cooking-book
Disbelief	00:01:37.000 – 00:01:38.010	Hmm
Ironic smile	00:01:43.010 – 00:01:44.000	he put cooking-book aside
Disbelief	00:01:57.000 – 00:01:58.000	desperate scene

Anna8

Labels	TC	Triggers
Anger	00:00:14.540 - 00:00:16.510	Men are mean animals
Satisfaction	00:00:19.480 - 00:00:21.010	that's right
Disbelief	00:00:24.990 - 00:00:26.510	also mean
Satisfaction	00:00:34.500 - 00:00:36.500	chocolate, full of nuts, almonds and all other things
Satisfaction	00:00:40.010 - 00:00:41.000	asked doubtfully
Disbelief	00:01:40.000 - 00:01:41.990	disability
Astonishment	00:02:00.000 - 00:02:00.990	fascinated
Satisfaction	00:02:11.010 - 00:02:13.000	terrible afraid of mom
Disbelief	00:02:28.000 - 00:02:29.010	frightened
Surprise	00:02:31.000 - 00:02:31.990	Chicken-Pox
Satisfaction	00:02:41.990 - 00:02:43.000	very surprised
Satisfaction	00:03:04.990 - 00:03:06.010	joyfully
Ironic smile	00:03:07.020 - 00:03:08.010	Aspazja from Millet

Anna9

Labels	TC	Triggers
Astonishment	00:00:02.410 - 00:00:03.610	what
Disbelief	00:00:03.990 - 00:00:05.510	this time
Ironic Smile	00:00:06.000 - 00:00:07.000	nothing unusual
Satisfaction	00:00:08.000 - 00:00:09.000	delicious
Understanding	00:00:11.000 - 00:00:12.000	birthday-party
Surprise	00:01:03.510 - 00:01:04.500	bubble
Disbelief	00:01:32.470 - 00:01:33.560	this mass!
Satisfaction	00:02:11.990 - 00:02:13.000	succeed

Anna10

Labels	TC	Triggers
Satisfaction	00:00:12.480 - 00:00:13.560	Colors of earth
Satisfaction	00:00:16.990 - 00:00:18.000	grumbled
Astonishment	00:00:47.000 - 00:00:48.000	in these clothes
Disbelief	00:00:56.440 - 00:00:57.530	more human
Ironic smile	00:02:03.010 - 00:02:05.980	full of style
Understanding	00:02:12.000 - 00:02:13.010	look like Clark Gable
Happiness	00:02:28.010 - 00:02:30.000	so, don't say anything

Annotator 3

Anna1

Labels	TC	Trigger
Disbelief	00:00:05.310 - 00:00:06.330	My name is..
Satisfaction	00:00:13.330 - 00:00:13.810	Borejko.
Surprise	00:00:18.320 - 00:00:20.110	This door?
Satisfaction	00:00:35.180 - 00:00:38.010	..to shut it. Oh, I understand
Understand	00:00:40.300 - 00:00:42.090	It blows from downstairs
Disbelief	00:00:45.690 - 00:00:46.610	It is possible
Happiness	00:00:49.400 - 00:00:51.750	so, have you also noticed it?
Disbelief	00:01:04.290 - 00:01:05.280	Interesting that...
Surprise	00:01:17.290 - 00:01:19.700	mysterious
Sadness	00:01:23.210 - 00:01:25.900	Well, so please remember this door.
Surprise	00:01:44.140 - 00:01:46.360	No?
Disbelief	00:01:47.600 - 00:01:48.630	cultured neighbor
Understand	00:01:52.090 - 00:01:54.440	Why aren't you?
Surprise	00:02:05.170 - 00:02:06.310	Bachelors..
Understand	00:02:06.840 - 00:02:08.570	banging doors
Disbelief	00:02:08.670 - 00:02:09.580	knocking
Disbelief	00:02:10.360 - 00:02:12.960	If you don't belief me, please check it
Disbelief	00:02:20.530 - 00:02:21.270	She grasped
Satisfaction	00:02:31.100 - 00:02:35.500	We will try to consider your wishes
Happiness	00:02:47.950 - 00:02:49.210	Goodnight

Anna2

Labels	TC	Trigger
Disbelief	00:00:09.030 - 00:00:09.950	back home..
Disbelief	00:00:10.160 - 00:00:12.320	Father was obstinately refusing
Disbelief	00:00:14.900 - 00:00:16.100	They told you..
Sadness	00:00:24.250 - 00:00:25.420	troubled and very sad
Disgust	00:00:26.960 - 00:00:27.810	lips of a sulked
Disbelief	00:00:39.390 - 00:00:40.250	Go alone
Ironic smile	00:00:41.280 - 00:00:44.010	what those doctors may find
Disbelief	00:00:54.600 - 00:00:55.650	passing them
Disgust	00:01:05.120 - 00:01:06.300	He must find pleasure
Surprise	00:01:18.630 - 00:01:19.240	when they..
Understanding	00:01:25.950 - 00:01:27.500	he will be here anole
Astonishment	00:01:29.200 - 00:01:29.870	I beg you
Astonishment	00:01:37.240 - 00:01:38.420	unusual for..
Understanding	00:01:42.680 - 00:01:43.900	That was the truth

Anna 3

Labels	TC	Trigger
Astonishment	00:00:05.330 - 00:00:06.330	You called yea?
Disbelief	00:00:14.450 - 00:00:15.140	What was it..
Surprise	00:00:17.300 - 00:00:18.090	ulcer on the..
Astonishment	00:00:26.760 - 00:00:28.060	she drank seed flax
Astonishment	00:00:38.760 - 00:00:39.960	Kowalik told me
Disgust	00:00:40.340 - 00:00:42.570	had to cut out half her stomach
Disbelief	00:00:47.540 - 00:00:48.750	How long will she stay?
Surprise	00:00:56.830 - 00:00:57.830	With everything
Understanding	00:01:00.830 - 00:01:02.040	no problem
Disbelief	00:01:04.620 - 00:01:05.310	what?
Satisfaction	00:01:10.060 - 00:01:10.860	forget about school

Anna4

Labels	TC	Trigger
Satisfaction	00:00:07.390 - 00:00:09.490	She looked happy
Satisfaction	00:00:14.910 - 00:00:15.570	Yes
Disbelief	00:00:17.230 - 00:00:18.060	girl answered
Disbelief	00:00:23.010 - 00:00:24.300	I think it's your sister
Disbelief	00:00:40.380 - 00:00:41.060	what's going on?
Satisfaction	00:00:42.220 - 00:00:44.150	don't get irritated ok?
Disbelief	00:00:47.940 - 00:00:48.740	Found what?
Satisfaction	00:00:54.300 - 00:00:55.610	frightening me
ironic smile	00:00:56.060 - 00:00:57.280	want to talk nonsense
Surprise	00:01:04.970 - 00:01:06.150	don’t you sleep yet?
ironic smile	00:01:10.310 - 00:01:12.100	darling, you can have three chickens..
Happiness	00:01:12.320 - 00:01:15.960	You can have a camel
Anger	00:01:47.010 - 00:01:47.890	talking
Disbelief	00:01:47.930 - 00:01:48.890	chicken-pox?
Astonishment	00:01:48.940 - 00:01:50.500	chicken-pox?
Satisfaction	00:01:51.680 - 00:01:52.670	chicken
Astonishment	00:02:00.300 - 00:02:02.090	"volatile virus"
Understanding	00:02:04.470 - 00:02:05.430	Oh Pulpa
Surprise	00:02:06.770 - 00:02:08.040	but in another place
Disbelief	00:02:08.100 - 00:02:08.970	Pap said
ironic smile	00:02:12.110 - 00:02:13.260	little bit different
Surprise	00:02:17.770 - 00:02:18.760	do you have a fever?
ironic smile	00:02:24.400 - 00:02:25.320	apartment thumbing
Disbelief	00:02:28.380 - 00:02:30.400	so persistently
Disbelief	00:02:34.240 - 00:02:38.230	What do you mean! what do you mean!

Anna 5

Labels	TC	Trigger
Disbelief	00:00:06.300 - 00:00:07.780	dragging to the bathroom
Satisfaction	00:00:10.820 - 00:00:12.480	she groaned
Surprise	00:00:14.480 - 00:00:15.600	Pulpa,,
Surprise	00:00:20.970 - 00:00:23.460	what happened
Ironic smile	00:00:23.480 - 00:00:25.290	could not sleep
Disbelief	00:00:28.110 - 00:00:29.090	did not close my eyes
Ironic smile	00:00:32.750 - 00:00:34.810	pretend sleeplessness
Satisfaction	00:00:37.100 - 00:00:39.250	ok, listen Mrs.....
Surprise	00:00:40.930 - 00:00:43.880	has forgotten that she is threatened...
Disbelief	00:00:45.650 - 00:00:48.530	she is an awful woman..
ironic smile	00:00:52.780 - 00:00:53.810	had to get up from
Astonishment	00:01:00.110 - 00:01:03.850	you were at her place tonight?
Disbelief	00:01:07.410 - 00:01:08.540	terrible
Surprise	00:01:14.430 - 00:01:17.580	Ida was staring at Gabrysia
Disbelief	00:01:19.320 - 00:01:20.290	feel worse and
Understanding	00:01:23.360 - 00:01:24.190	every minute
Disbelief	00:01:38.420 - 00:01:46.110	I heard a strange noise,,,
Disgust	00:01:46.640 - 00:01:48.600	something crashed dully, and...
Understanding	00:01:57.530 - 00:01:58.930	all of it is true
Astonishment	00:02:00.200 - 00:02:02.110	she isn't crazy?
Disbelief	00:02:02.750 - 00:02:04.020	watch your manners
Disbelief	00:02:04.140 - 00:02:04.960	right in all aspects
Anger	00:02:05.880 - 00:02:07.850	she can hear everything
Disbelief	00:02:22.630 - 00:02:24.430	us making noises
Astonishment	00:02:26.930 - 00:02:27.940	something was going
surprise	00:02:33.630 - 00:02:34.840	at midnight

Anna 6

Labels	TC	Trigger
Sadness	00:00:10.860 - 00:00:12.310	seemingly
ironic smile	00:00:13.850 - 00:00:15.010	blushed ridiculously
Happiness	00:00:17.310 - 00:00:18.590	that's very nice
Surprise	00:00:26.540 - 00:00:27.590	like to get it?
Satisfaction	00:00:28.010 - 00:00:29.970	oh yes gladly
Disbelief	00:00:35.990 - 00:00:37.170	that we have a chick
Surprise	00:01:02.030 - 00:01:03.430	really?
Satisfaction	00:01:09.090 - 00:01:10.850	water
Disbelief	00:01:12.440 - 00:01:14.130	I’m sick, I’m dying
Disbelief	00:01:16.830 - 00:01:17.850	faint out of fear
Surprise	00:01:26.290 - 00:01:27.000	..please..
Astonishment	00:01:31.980 - 00:01:32.940	Your fiancé
Sadness	00:01:35.060 - 00:01:35.720	well I understand
Disgust	00:01:42.750 - 00:01:44.120	beat res-haired severally

Anna 7

Labels	TC	Trigger
Satisfaction	00:00:05.430 - 00:00:06.060	refrigerator
Astonishment	00:00:11.100 - 00:00:11.920	organize it somehow
Astonishment	00:00:12.500 - 00:00:13.180	go back to
Disbelief	00:00:18.340 - 00:00:19.160	what a pity
Astonishment	00:00:23.790 - 00:00:24.930	laying in bed already
Disbelief	00:00:33.790 - 00:00:35.190	did not hear me
Disbelief	00:00:38.790 - 00:00:39.550	I heard you
Ironic smile	00:00:40.010 - 00:00:42.000	my darling
Disbelief	00:00:56.730 - 00:00:57.720	to take care about.
Understanding	00:00:59.000 - 00:01:00.000	father struggled with his urge
Satisfaction	00:01:05.820 - 00:01:06.830	think I can cook
Satisfaction	00:01:25.640 - 00:01:27.540	father awaked
Disbelief	00:01:37.000 - 00:01:38.010	Hmm
Ironic smile	00:01:43.010 - 00:01:44.000	he put cooking-book aside
Disbelief	00:01:57.000 - 00:01:58.000	desperate scene

Anna 8

Labels	TC	Trigger
disbelief	00:00:06.460 - 00:00:07.800	ida, don’t worry
disbelief	00:00:11.540 - 00:00:12.770	hemstitch?
sadness	00:00:14.830 - 00:00:16.510	depression
satisfaction	00:00:18.600 - 00:00:19.830	oh, that's right
disbelief	00:00:25.030 - 00:00:26.240	and Pyziak, he also..
satisfaction	00:00:29.470 - 00:00:30.590	life will be just a little
ironic smile	00:00:34.500 - 00:00:35.960	dripping with chocolate
ironic smile	00:00:36.540 - 00:00:38.850	all other stuff
surprise	00:00:39.490 - 00:00:41.110	Ida asked doubtfully
disbelief	00:00:45.460 - 00:00:46.570	well, in any case
satisfaction	00:00:47.440 - 00:00:48.300	already today
disbelief	00:01:14.400 - 00:01:15.530	this is Klaudius
surprise	00:01:24.640 - 00:01:25.690	he went out
sadness	00:01:39.980 - 00:01:41.130	some sort of disability
satisfaction	00:02:11.010 - 00:02:12.580	remember Walus
disbelief	00:02:28.590 - 00:02:29.780	rascal
surprise	00:02:31.080 - 00:02:32.510	still to regain

Anna 9

Labels	TC	Trigger
satisfaction	00:00:02.270 - 00:00:03.690	Gabrysia speaking
disbelief	00:00:03.940 - 00:00:05.170	what happened this time
happiness	00:00:06.780 - 00:00:08.840	warmed up dinner again
surprise	00:00:10.480 - 00:00:12.170	make a fancy cake
disbelief	00:00:13.420 - 00:00:14.870	oh my God, Grabrysia
disbelief	00:00:17.410 - 00:00:18.230	clumsy hands
satisfaction	00:00:21.780 - 00:00:23.280	play it with hands
disbelief	00:00:49.110 - 00:00:51.170	impossible to buy
understanding	00:00:54.580 - 00:00:55.850	it will be alright for your
surprise	00:01:01.690 - 00:01:02.880	clear?
disbelief	00:01:04.420 - 00:01:05.590	bubble?
satisfaction	00:01:14.120 - 00:01:15.310	egg white?
disbelief	00:01:18.260 - 00:01:19.180	be more patient
understanding	00:01:20.780 - 00:01:22.580	oh. Why are you so..
disbelief	00:01:32.240 - 00:01:34.000	this mass!
understanding	00:01:35.230 - 00:01:36.180	oh
ironic smile	00:01:52.910 - 00:01:53.970	...and pour it into a
Disbelief	00:01:57.820 - 00:01:59.110	end of the receipt
satisfaction	00:02:10.970 - 00:02:13.170	decided to be a feminine

Anna 10

Labels	TC	Trigger
Astonishment	00:00:12.080 - 00:00:13.420	colors of earth
satisfaction	00:00:16.690 - 00:00:17.740	in my arse
surprise	00:00:45.540 - 00:00:46.650	why should I..
astonishment	00:00:46.870 - 00:00:49.630	bursted out laughing
surprise	00:00:51.150 - 00:00:51.730	appropriate
sadness	00:00:55.890 - 00:00:56.970	may academy
ironic smile	00:01:01.060 - 00:01:01.930	nodded
disbelief	00:01:12.770 - 00:01:14.250	why?
ironic smile	00:02:03.240 - 00:02:04.350	imitation
surprise	00:02:09.820 - 00:02:10.920	one week
happiness	00:02:11.150 - 00:02:12.520	put pap's hat on my head
surprise	00:02:21.590 - 00:02:22.120	hat?
satisfaction	00:02:28.430 - 00:02:30.590	don’t say..
ironic smile	00:02:45.840 - 00:02:47.410	powder sugar

Recording Protocol
For the development of a Multimodal Database

Introduction

The creation of a new multimodal database is a lot of work. Normally it takes years to fill the database with enough representative subjects in order to train systems that should be generalizable enough.

This booklet will briefly describe the recording protocol in order to capture the expressions expressed by the participants. Besides that, this booklet covers the set of emotions to be used, the set up of the room, the search for subjects, the procedure used to record the subjects and the equipment used. Finally there is suggestion of how to validate and annotate the recordings. The used stories are also included in this booklet, but are all in Dutch. This is done because we have a Dutch speech recognizer.

This audio-visual emotion database should be used as a reference database for testing and evaluating video, audio or joint audio-visual emotion recognition algorithms. Additional uses may include the evaluation of algorithms performing other multimodal signal processing tasks, such as multimodal person identification or audio-visual speech recognition.

Emotional Content

The emotional content of the database should include at least the six basic emotions described by Ekman, some more mixed emotions and a neutral recording of each subject must be present. Emotions can be expressed with occlusion, like glasses or facial hair. The occlusion should not prevent good recognition from the profile view. Emotions to be expressed are summed up in Table 1. The Dutch translations are given as well.

Table 1 The different emotions captured

	English	Dutch
1	Admiration	Bewondering
2	Amusement	Amusement
3	Anger	Boos
4	Boredom	Verveling
5	Contempt	Minachting
7	Desired	Verlangen
6	Disappointment	Teleurstelling
8	Disgust	Walging
9	Dislike	Afkeer
10	Dissatisfaction	Ontevredenheid
11	Fascination	Geboeid
12	Fear	Bang
13	Furious	Kwaad
14	Happiness	Blijdschap
15	Indignation	Verontwaardiging
16	Inspired	Inspiratie
17	Interest	Interesse
18	Pleasant surprise	Aangenaam verrast
19	Sadness	Verdrietig
20	Satisfaction	Tevredenheid
21	Unpleasant surprise	Onaangenaam verrast

The environment conditions

The experiment is done in a closed room with good lighting conditions. Good lighting conditions mean that there is enough diffuse light to leave no shadows on the participants face. The camera is focused on the participant and the height of the camera is just right for lip reading. A consequence of this is that the height of the camera must change for every participant in order to get a perfect frontal view. The background behind the participant is covered with a dark color, preferably dark blue or black.

The room setting

In the room there is a chair on which the participants will take a seat. Right in front of this chair there is a camera, about one or two meters from the participant. A mirror is placed on the left side of the participant, giving the camera a good silhouette view of the participants face. The chair is placed not too far from the wall, so the background can be covered by a large, dark colored, piece of cloth. An overview of this situation is given in Figure 1 and a map in Figure 2.

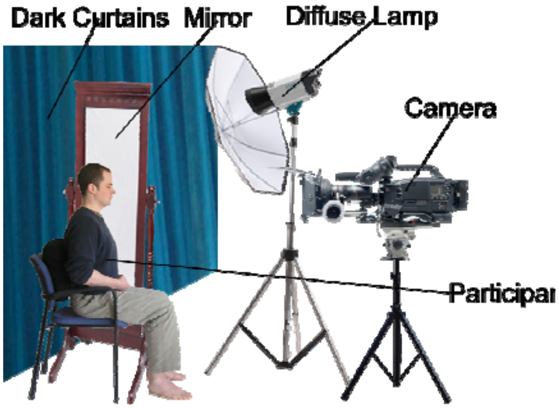


Figure 1 Overview of the room set up

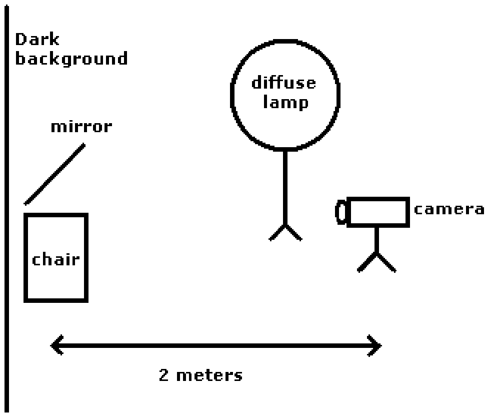


Figure 2 Map of the room

The participants

The search for participants is a task that should be taken seriously. The distribution of the subjects should be a mirror of the population and there are more things to consider like facial hair, glasses or hair that falls in front of the face. All these aspects will make sure that you have a good distribution of different persons in your database. Age is also important, for the training of systems is might even be good to have samples of the same subject, but with different ages involved (this process may take years).

Be sure to recruit only native speakers, as non-native speakers tend to think about what they are going to say so that the expressed emotion is less likely to be 'spontaneous' or totally unnatural. In our case all speakers must be native Dutch speakers.

The procedure of the experiment

To obtain the subject's emotional expressions the experimenter should instruct the subject to perform the pure facial and vocal expression. The procedure of the experiment is as follows.

1. The subject is told about the short stories to which he or she has to react to. The corresponding emotions are explained beforehand, so the participant knows which emotion to express.
2. The participant gets a list with responses corresponding to the expected expressed emotion.
3. The participant is asked to sit before the frontal camera, while the mirror is to the left of the participant. The distance between the camera and the participant is about two meters.
4. The participant listens to or reads the short story and is asked to imagine being in this situation.
5. The experimenter gives the order to capture the emotional expressions of the participant.
6. The participant is then asked to react with the five pre-defined sentences.
7. If the participant's performance is not ideal, repeat step 4 to 6
8. After the experiment is done, thank the participant for the cooperation.

Repeat step 4 to 7 for the next designated facial expression, until the participant has displayed all of the facial expressions.

The subjects are asked to put as much expressiveness as possible, producing a message that contains only the emotion to be elicited. When the subject obviously don't know how an emotion should be expressed, the experimenter should suggest a way of expressing it, based on their knowledge of the way the emotion is generally expressed.

Validation

There can be a discrepancy between the perceived and experienced emotion: people may not always express their (true) emotions, especially when they are in conversation and obey the unwritten conversational rules. An option would be to let the subjects annotate their own emotional expressions (self annotations) and compare these with the annotated emotions as perceived by other subjects.

Validation Methods

The recorded videos must be annotated to what the majority of observers agree upon. Therefore the videos should be validated by as many different subjects as possible, as well as the participants themselves and experts in the field of emotion recognition.

Stories (in Dutch)

Admiration

“Je loopt samen met een vriend/vriendin door een dure winkelstraat in Amsterdam en ziet in de etalage een jas hangen die je altijd al had willen hebben. Je droomt over wat je zou doen als je het geld had om deze jas te kopen. Je gaat voor de etalage staan en denkt...”

Reactions

R1: Oooohhh..	o
R2: Dat ziet er goed uit.	dat zit @r Gut Y+t
R3: Die zou ik graag willen hebben.	di zA+ Ik GraG wIl@ hEb@
R4: Was die maar van mij.	wAs di mar van mE+
R5: Zodra ik mijn geld heb, is die jas van mij.	zodra Ik mE+n (m@n) GElt hEp Is ti jAs fAn mE+

Amusement

“Jij loopt met een vriendin op straat langs een kiosk. Opeens zie jij de foto van je favoriete zanger op de cover van een tijdschrift. Onder de foto staat dat hij binnenkort in een film van een beroemde regisseur de hoofdrol zal spelen. Jij vindt het werk van die regisseur ook heel goed. O, wat wil je die film graag zien.”

Reactions

R1: Dat zal een goeie film worden.	dAt zal en Guj@ film wOrd@
R2: Die film wil ik zeker zien.	di film wIl Ik zek@r zin
R3: Hij maakt altijd goede films.	hE+ makt AltE+t Gud@ films
R4: Wanneer gaat deze film in première?	wAner Gat dez@ fIlm In pr@miE:r@
R5: Dat wordt kijken geblazen!	dAt wOrt kE+k@ G@blaz@

Anger (Surprise)

“Je zit thuis op de bank lekker televisie te kijken als er plots wordt aangebeld. Je loopt naar de deur en doet open. Voor je staat een vreemde vrouw die nogal nerveus over komt. Ineens barst ze in tranen uit en vertelt ze dat zij tijdens het parkeren jouw auto heeft aangereden! Je dacht dat de auto veilig in de garage stond, maar langzaam herinner je dat je vanmiddag nog snel even bent wezen winkelen en de auto voor je huis hebt laten staan.”

Reactions

R1: Wat??? Nee, nee dat kan niet!	wAt ne ne dAt kAnit
R2: Dat gaat je geld kosten!	dAt Gat j@ GElt kOst@
R3: Weet je zeker dat het mijn auto is!?	wet j@ zek@r dAt @t mE+n A+to Is
R4: Zie je wel, vrouwen kunnen niet rijden!	zi j@ wEl vrA+Yn k@n@ nit rE+d@
R5: Niet mìjn auto? Het is niet waar!	nit mE+n A+to hEt Is nit war

Anger

“Je vriend(in) neemt je mee naar een duur restaurant om jullie 5 jarig samen zijn te vieren. Samen met je vriend(in) genieten jullie van de avond en het restaurant. Nog geen idee wat jullie zullen bestellen komt de ober vragen of jullie vast een aperitiefje willen. Je stemt in en laat vast een stokbrood met kruidenboter komen en twee biertjes. Het brood is fantastisch en zulke lekkere kruidenboter heb je nog nooit geproefd, je kan niet wachten tot de soep er is. Als de soep eenmaal arriveert laat de ober per ongeluk de gloeiend hete soep over je heen vallen. Je hele avond is verpest en je humeur wordt zeker niet beter.”

Reactions

R1: AUW! Klootzak!	A+ klotsAk
R2: Pas op met die hete soep!	pAs op mEt di het@ sup
R3: Hier kom ik dus nooit meer!	hir kOm Ik d@s nojt mer
R4: Nee! Mijn nieuwe broek verpest!	ne mE+n niw@ bruk v@rpEst
R5: Kom, we gaan!	kOm w@ gan

Boredom

"Je zit alleen op de bank op je vrije zaterdagmiddag. Vroeger ging je nog wel een naar het sportveld, maar tegenwoordig niet meer. Je hebt eigenlijk weinig te doen en je verveelt je een beetje. Je moeder komt binnen en vraagt 'Verveel je je?'. Waarop jij antwoord:"

Reactions

R1: Mwah...	mwAh
R2: Laat me maar...	lat m@ mar
R3: Huh?	h@
R4: Hmmmmmm. Kweenie.	hm kweni
R5: Neuh..	n@

Contempt

"Van een goede vriend hoor je dat je vervelende buurman, die altijd loopt te roddelen over je, van de trap gevallen is. Hij heeft zijn neus gebroken."

Reactions

R1: Veel slechter kan het niet worden.	vel sleGt@r kAn @t ni wOrd@
R2: O, het is Jaap maar.	o hEt Is jap mar
R3: O die, ja dat dacht ik wel.	o di ja dAt dAGt Ik wEl
R4: Met zijn neus recht in de boter!	mEt z@n_n2s In d@ bot@r
R5: Net goed.	nEt Gut

Desired

"Je loopt langs een elektronica winkel en ziet in de etalage een prachtige 50inch breedbeeld LCD televisie staan. Zo een heb je altijd al willen hebben, maar hij is natuurlijk veel te duur voor je. Je blijft stil staan en kijkt er nog eens goed naar."

Reactions

R1: Aaaahhh..	a
R2: Dat ziet er goed uit!	dAt zit@r Gut Y+t
R3: Die/Dat zou ik graag willen hebben.	di zA+ Ik GraG wIl@ hEb@
R4: Had ik er ook maar een.	hAt Ik Er ok mar en
R5: Wóóów.	wow

Disappointment

"Je laatste eindexamen was zeker twee weken geleden en de cijfers moeten reeds bekend zijn. Vanmiddag zal je worden gebeld en iemand zal je vertellen wat de uitslag is en dan kan je gaan genieten van je zomervakantie. Je hebt zo hard je best gedaan, maar je weet niet zeker of je geslaagd bent. Als de telefoon over gaat ben je als eerste bij de telefoon om hem op te nemen. Je hoort een stem aan de andere kant zeggen dat je gezakt bent."

Reactions

R1: Ik ben gezakt.	Ik bEn G@zAkt
R2: Jammer, volgende keer beter.	jAm@r vOlG@nd@ ker bet@r
R3: Dat had ik niet verwacht.	Dat hAt Ik nit v@rwAGt
R4: En wat moet ik nu?	En wAt mut Ik ny
R5: Nee hè!	ne hE

Disgust

"Vrolijk loop je over straat op weg naar je nieuwe baan. In de verte zie je een vage bekende. Je kijkt nog eens goed en probeert te herinneren wie het is. Als je even niet op let stap je in een grote, vers gelegde, hoop poep. Je goede schoenen zitten er volledig onder, zo kan je toch niet verschijnen op je sollicitatiegesprek. Je baalt en kijkt naar de viezigheid beneden."

Reactions

R1: Ieeuuwww, poep.	i@w pup
R2: Wat is dat vies.	wAt Is dAt vis
R3: Hè, bah!	hE bA
R4: Niet nu.	nit ny
R5: Gatver, mijn schoenen vies.	GAtf@r mE+n sGun@ vis

Dislike

"Je moeder heeft voor je verjaardag een, volgens haar, leuk cadeau gekocht. Ze neemt je mee naar buiten en laat een gloed nieuwe tent zien. Je denk 'ze weet toch dat ik niet van kamperen hou.' En kijkt haar een beetje teleurgesteld aan."

Reactions

R1: Een groene?!	@n Grun@
R2: Je weet toch dat ik niet van kamperen hou.	j@ wet tOG dAt Ik nit vAn kAmper@ hA+
R3: Wat moet ik hier nu weer mee?	wAt mut Ik hir ny wer me
R4: Dit vindt ik toch niet leuk.	dIt vInt Ik tOG nit l2k
R5: Ik vind hem niet eens mooi!	Ik vInt hem nit ens moj

Dissatisfaction

"Die mooie jas die je laatst had gekocht is stuk gegaan en je gaat er mee terug naar de winkel. Je wilt je geld terug, maar de mevrouw achter de kassa zegt dat je wel een tegoed bon kan krijgen, maar niet je geld terug. Je bent het er eigenlijk niet mee eens, maar accepteert de bon toch.

Reactions

R1: Dan neem ik deze maar.	dAn nem Ik dez@ mar
R2: Ach, het is beter dan niets	AG hEt Is bet@r dAn nits
R3: Toch liever mijn geld gehad.	tOG liv@r m@n GElt G@hAt
R4: Jammer, toch bedankt.	jAm@r tOG b@dAnkt
R5: Ik snap niet waarom ik mijn geld niet terug krijg	Ik snap nit warOm Ik mE+n GElt nit_tr@G krE+G

Fascination

"Je kijkt naar een documentaire over het heelal. Je merkt hoe onvoorstelbaar groot het heelal is en hoe klein onze kleine aarde. Na de afloop van het programma denk jij… "

Reactions

R1: Wow wat een verhaal.	wA+w wAt @n v@rhal
R2: Dat wist ik allemaal niet.	dAt wist Ik Al@mal nit
R3: Het was echt een interessant verhaal.	het was EGt en Int@r@sAnt vErhal
R4: Het is echt indrukwekkend.	het Is EGt Indr@kwEk@nt
R5: ohh, fascinerend!	o fasiner@nt

Fear

"Je bent allen thuis en ligt in bed. Je slaapkamer is boven in het huis en jij bent de enige persoon die thuis is. Plotseling hoor je een geluid beneden. Je weet zeker dat er niemand zou thuiskomen vanacht. Je houdt je stil en weet zeker dat er iemand beneden is. Waarschijnlijk een dief, of zelfs een moordenaar! Je hoort hem de trap op lopen en je wordt heel erg bang."

Reactions

R1: Oh mijn God, er is iemand binnen!	o mE+n GOt Er Is imAnt bIn@
R2: Er komt iemand naar boven.	Er kOmt imAnt nar bov@
R3: Vermoordt me alsjeblieft niet...	vErmort m@ AlS@blift nit
R4: Help!	hElp
R5: Alsjeblieft, doe me niks!	AlS@blift du m@ nIks

Furius

"Je loopt in een onbekende stad en je hebt € 200 nodig om terug naar huis te komen. Er is maar één bank in deze stad. Je moet het geld echt vandaag nog hebben. De pinautomaat buiten is buiten gebruik en je kent verder niemand hier. De bank is tot vijf uur open en je stapt om kwart over vier binnen. Je ziet een enorme rij voor de balie staan en gaat achteraan staan. Na drie kwartier wachten als je eindelijk vooraan staat, de bankbediende vraagt of je morgen terug kunt komen, omdat hij nu zijn koffiepauze gaat houden voordat hij de bank sluit en naar huis gaat. Je legt je situatie uit, de bank moet nog zeker 15 minuten open zijn en je hebt het geld echt vandaag nodig, maar de man wil er niets van weten. Hij blijft maar herhalen dat het niet zijn probleem is en nu een koffiepauze gaat houden. Je bent woedend en zal niet thuis komen."

Reactions

R1: Wat??? Nee, nee, nee, luister! Ik moet dit geld vandaag hebben!	wAt ne ne lY+st@r Ik mut dIt GElt vAndaG heb@
R2: Die koffie kan me niets schelen, help mij liever!	di kofi kAn m@ nit sGel@ hElp mE+ liv@r
R3: Ik wil je baas spreken, NU!	Ik wIl j@ bas_sprek@ ny
R4: Is je koffie bELANGrijker dan mij helpen?	Is j@ kOfi b@lANrE+k@r dAn mE+ hElp@
R5: Je krijgt betaalt om te werken, niet om koffie te drinken!	j@ krE+Gt b@talt Om t@ wErk@ nit Om kOfi t@ drINk@

Happiness

“Je gaat een avond stappen in Holland Casino en besluit om vlak voordat je naar huis gaat nog even je laatste geld in de fruitautomaat te werpen. Vol verwachting blijf je wachten als het eerste kroontje verschijnt. Drie kroontjes is de jackpot, en de tweede valt. Je staat bijna op springen als je gestaag op het derde wiel wacht. Ja! Ook een kroon, je hebt de jackpot gewonnen en bent hartstikke blij.”

Reactions

R1: JAAAAAHHH!!!	ja
R2: Gewonnen!	G@wOn@
R3: Nu kan ik lekker op vakantie!	ny kAn Ik lEk@r Op vakAntsi
R4: O, dankjewel!	o dANkj@wEl
R5: Wat heerlijk.	wAt herl@k

Indignation

“Je loopt over straat naar een vriend. Het is best druk en je moet goed opletten, anders loop je nog tegen iemand aan. Als je telefoon gaat haal je deze uit je zak en let je even niet op. Je loopt tegen iemand aan die gestrekt op de grond valt. Als de man overeind komt en direct begint te schreeuwen weet je eerst niet hoe je moet reageren. Hij scheldt je uit voor van alles en nog wat.”

Reactions

R1: Euhh, sorry.	@ sOri
R2: Pardon, ik lette niet op.	pArdOn Ik lEt@ nit Op
R3: Rustig maar, u heeft niets gebroken hoor.	r@stIG mar y heft nits G@brok@ hor
R4: Nou zeg, ik deed het niet expres.	nA+ zEG Ik det hEt nit EksprEs
R5: Het was niet mijn bedoeling om u omver te lopen.	hEt wAs nit mE+n b@dulIN Om y OmvEr t@ lop@

Interest

“Op een feestje van je een vriendin, sta je met iemand te praten. Je kent hem niet zo goed, maar hij houd een mooi verhaal. Hij vertelt jou dat hij zweefvliegles geeft. Je wilde altijd al zweefvlieglessen nemen. Hij heeft een oranje lestoestel. Je denkt”

Reactions

R1: Ohh, wat leuk een oranje.	O wAt l2k en orAJ@
R2: Ik wilde altijd al zweefvliegles nemen.	Ik wIld@ AltE+t Al zwefvliGlEs nem@
	zwefvliG@ wIld@ Ik AltE+t nOG ens dun
R3: Zweefvliegen wilde ik altijd nog eens doen.	Gef jE+ EGt zwefvliGlEs
	kan j@ het mE+ ler@
R4: Geeft jij echt zweefvliegles?	
R5: Kan je het mij leren?	

Pleasant surprise

“Als je ‘s middags thuis aan het stofzuigen bent wordt er aangebeld. Je verwacht geen bezoek en hebt ook niemand aan horen komen. Als je open doet zie je daar Ron Brandsteder jr. Staan met een cameraploeg voor je deur. ‘Gefeliciteerd’ zegt hij, ‘U heeft 100.00 euro gewonnen’. Met stomheid geslagen sta je in de deuropening en kijkt hem aan.”

Reactions

R1: Heb ik gewonnen?	hEb Ik G@wOn@
R2: Voor mij?	vor mE+
R3: Eindelijk heb ìk geluk!	E+nd@l@k hEb Ik G@l@k
R4: Héél erg bedankt!	hel ErG b@dANkt
R5: Wahoo, Dat had ik nooit gedacht!	wahu dAt hAt Ik nOjt G@dAGt

Sadness

"Je komt thuis na een dag hard werken en ploft op de bank. Je wilt lekker ontspannen als plotseling de telefoon gaat. Het is je moeder, en ze klink nogal erg verontrustend. Ze verteld je dat je vader is overleden. Eerst wil je het niet geloven, maar na een tijdje dringt het tot je door wat dit voor je betekend. Als je terug denkt aan alle fijne momenten samen en dan bedenkt dat alles voortaan anders is voel je je erg verdrietig."

Reactions

R1: Het is niet waar, alsjeblieft.	hEt Is nit war AlS@blift
R2: NEEEEEE!	ne
R3: Ik snap het niet, hij was nog zo jong!	Ik snAp hEt nit hE+ wAs nOG zo jON
R4: Dit kan niet waar zijn.	dIt kAn_nit war zE+n
R5: Wat erg, wat moet ik nu.	wAt ErG wAt mut Ik ny

Satisfaction

"Op een warme zomerdag zit je op een terrasje en drink je een koud biertje. Terwijl je een slokje neemt geeft dit je een bevredigend gevoel. Wat is het leven toch goed."

Reactions

R1: Dit is echt lekker bier.	dit Is EGt lek@r bir
R2: Wat een perfecte zomerdag.	wAt @n pErfEkt@ zom@rdAG
R3: Dit voelt echt goed.	dit vult EGt gut
R4: Ik kan hier geen genoeg van krijgen.	Ik kAn hir Gen G@nuG vAn krE+g@
R5: Wat wil je nog meer?	wAt wIl j@ nOG mer

Unpleasant surprise

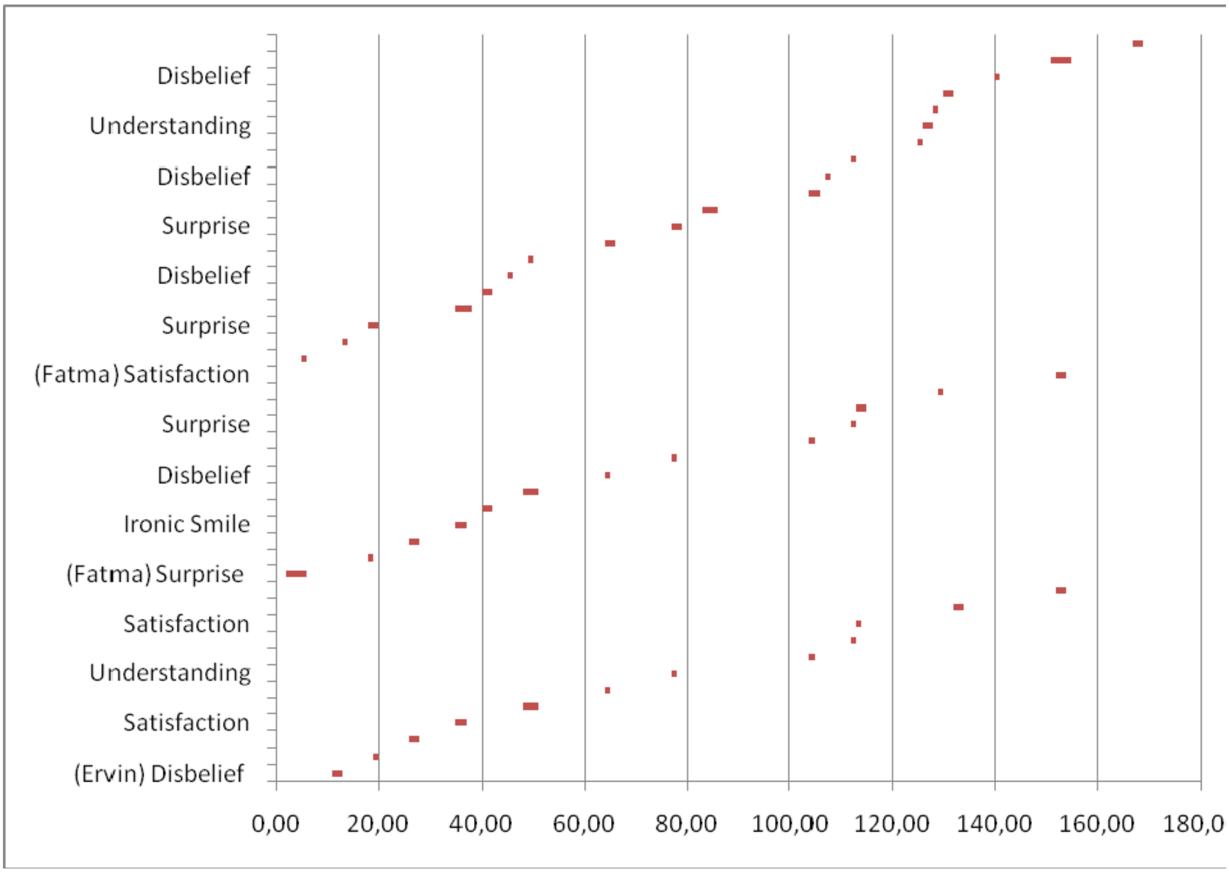
"Je partner zegt tegen je dat jullie eens serieus moeten praten over jullie relatie. Als jullie samen op de bank zitten zegt hij uit het niets tegen je dat hij homoseksueel is. Je bent erg verbaasd en had dit zeker niet verwacht."

Reactions

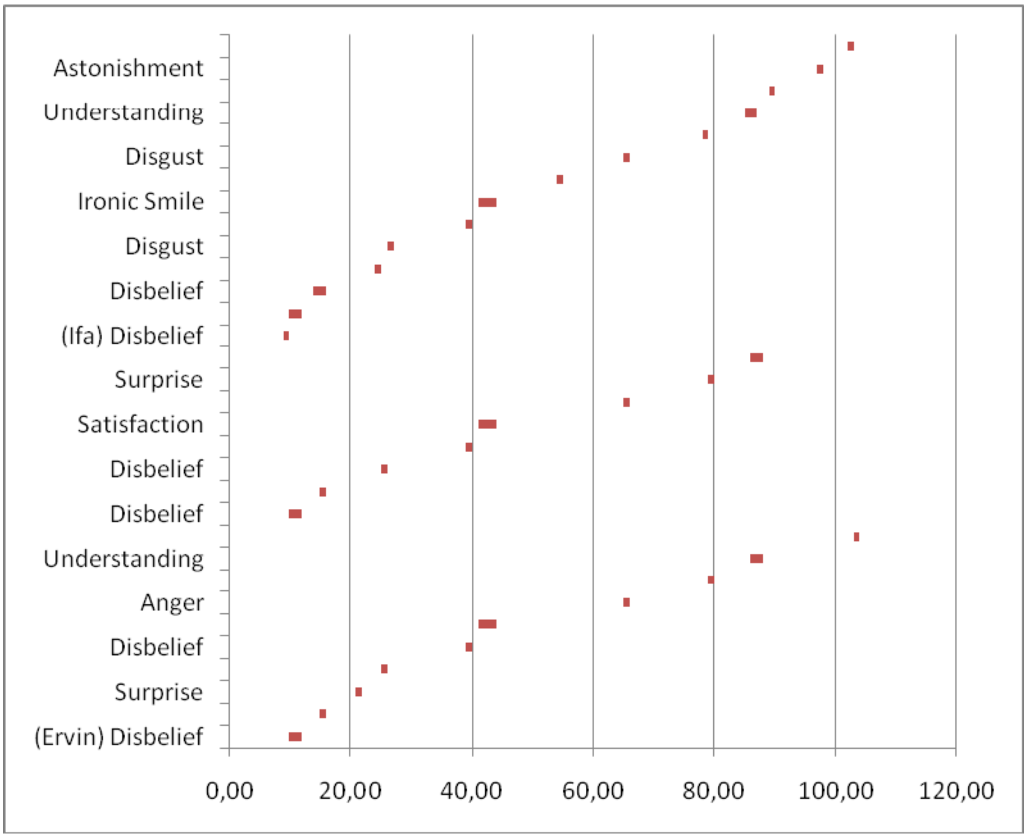
R1: Wat zeg je me nu!	wAt zEG j@ m@ ny
R2: Dat had ik niet verwacht!	dAt hAt Ik nit v@rwAGt
R3: Dit geloof je toch niet!	dIt g@lof j@ tOG nit
R4: Dat meen je niet!	dAt men j@ nit
R5: O mijn God, het is niet waar!	o mE+n GOt hEt Is nit war

D: Assigned labels by annotators

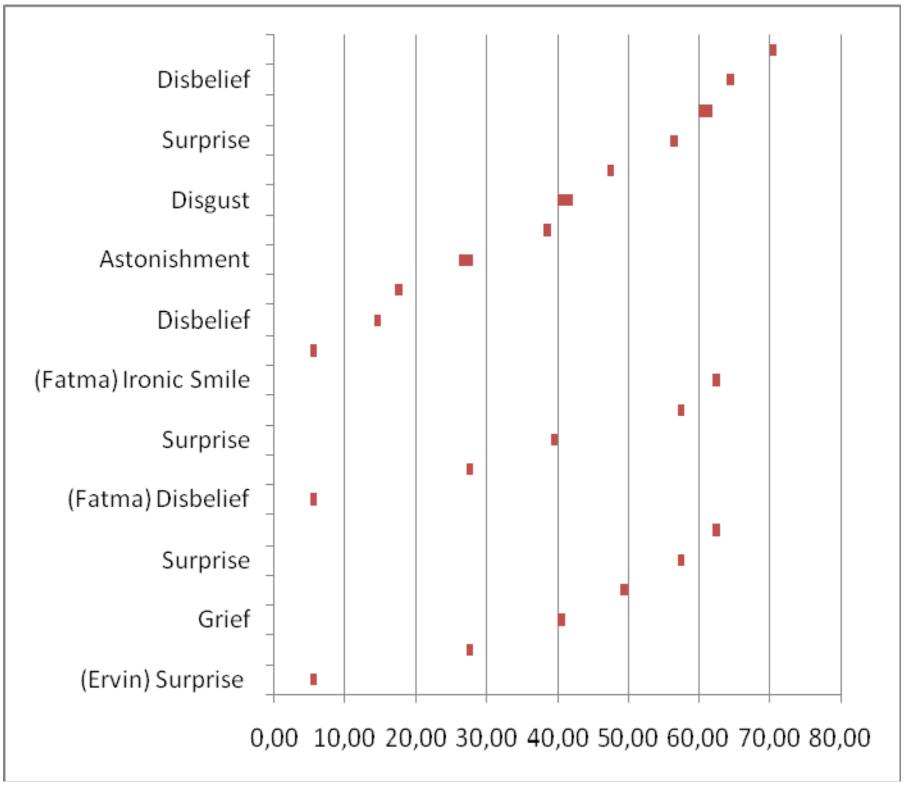
Anna 1



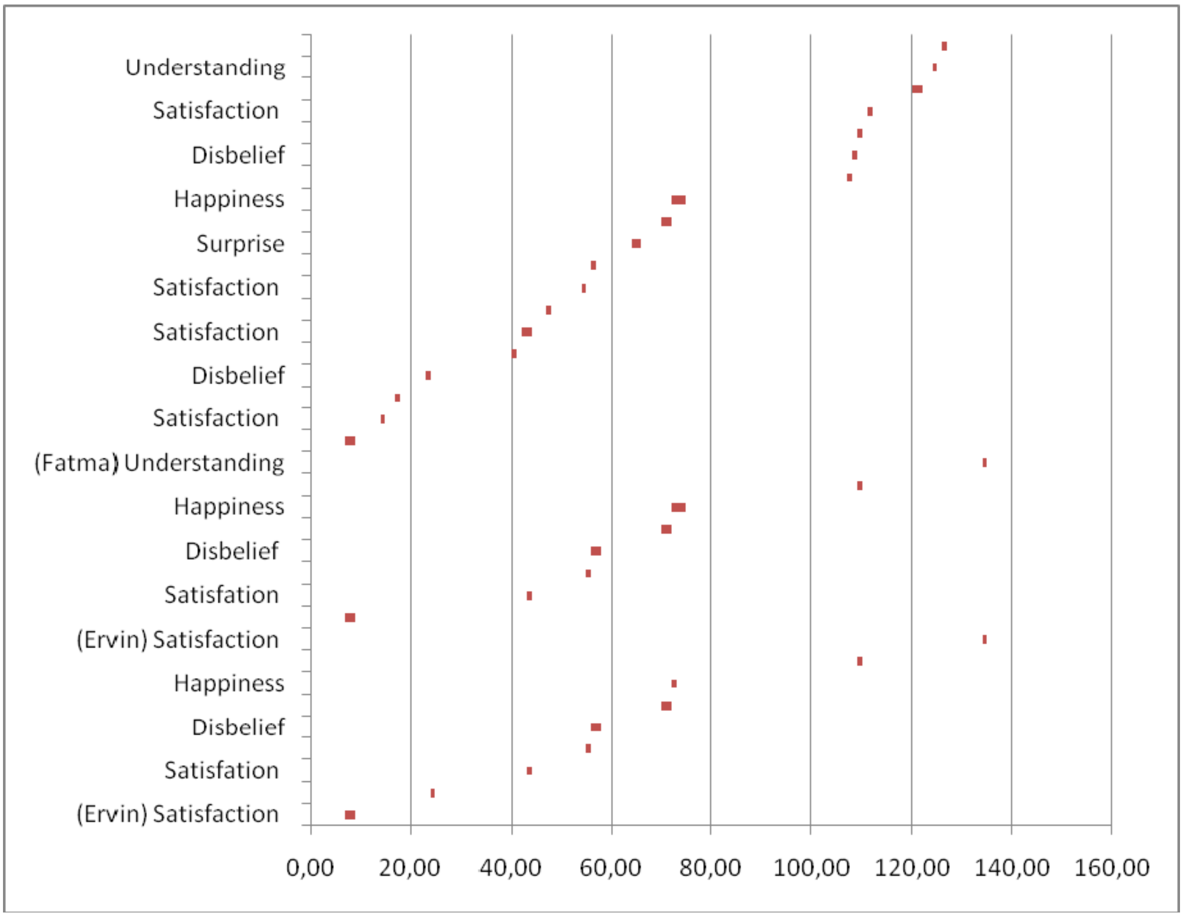
Anna 2



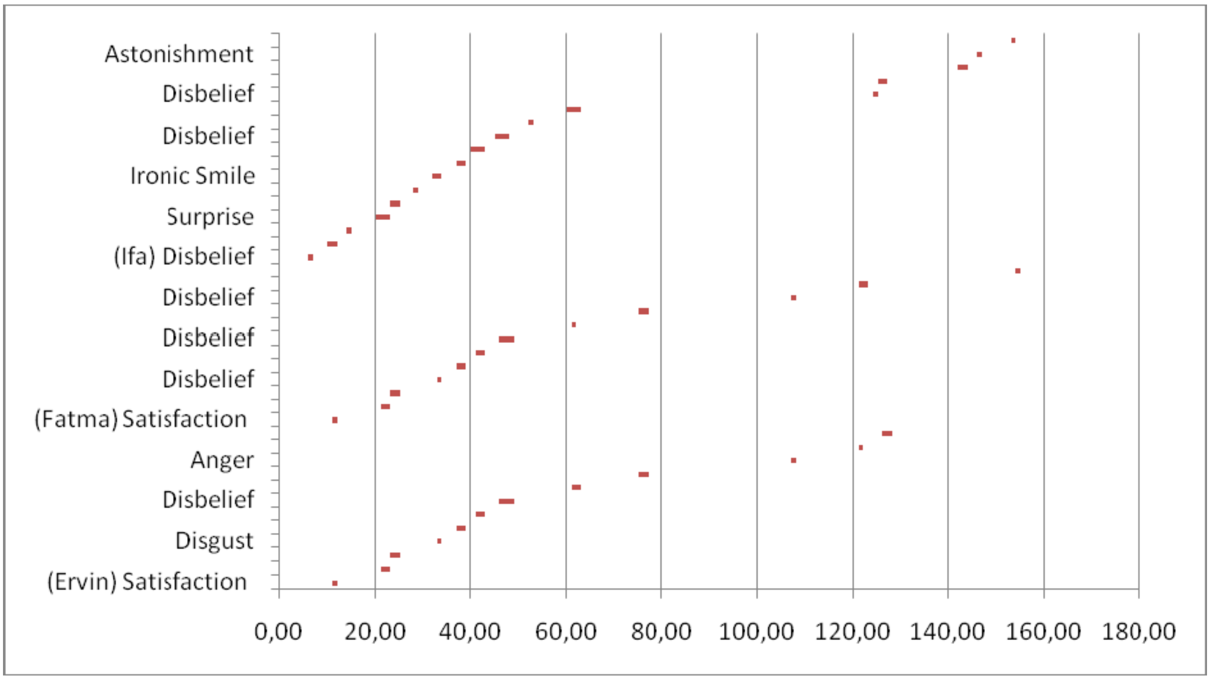
Anna 3


Disbelief
Surprise
Disgust
Astonishment
Disbelief
(Fatma) Ironic Smile
Surprise
(Fatma) Disbelief
Surprise
Grief
(Ervin) Surprise
0,00
10,00
20,00
30,00
40,00
50,00
60,00
70,00
80,00

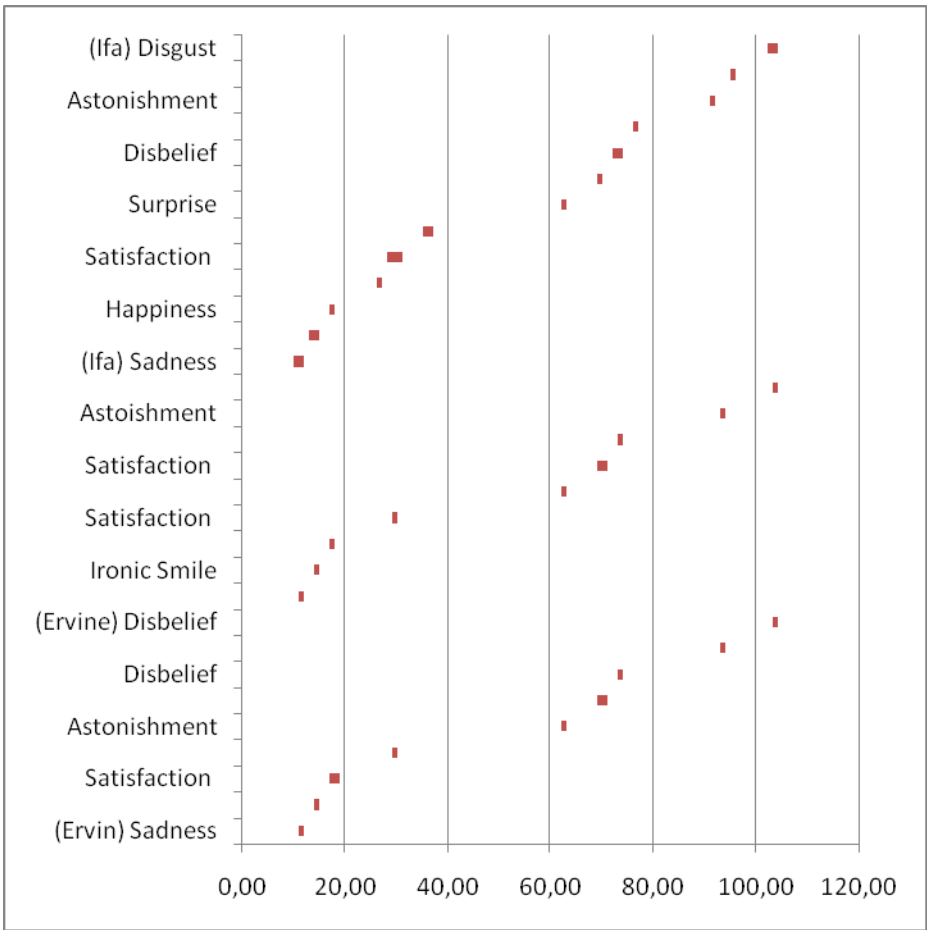
Anna 4


Understanding
Satisfaction
Disbelief
Happiness
Surprise
Satisfaction
Satisfaction
Disbelief
Satisfaction
(Fatma) Understanding
Happiness
Disbelief
Satisfation
(Ervin) Satisfaction
Happiness
Disbelief
Satisfation
(Ervin) Satisfaction
0,00
20,00
40,00
60,00
80,00
100,00
120,00
140,00
160,00

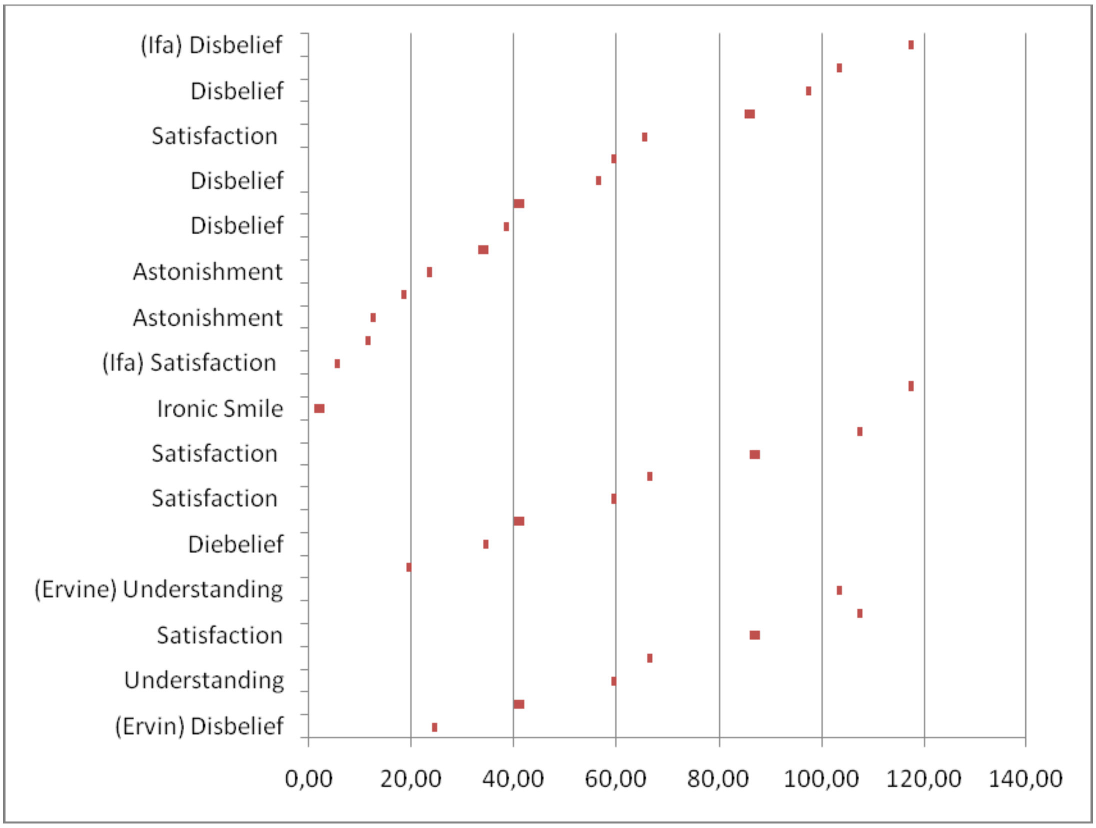
Anna 5


Astonishment
Disbelief
Disbelief
Ironic Smile
Surprise
(Ifa) Disbelief
Disbelief
Disbelief
Disbelief
(Fatma) Satisfaction
Anger
Disbelief
Disgust
(Ervin) Satisfaction
0,00
20,00
40,00
60,00
80,00
100,00
120,00
140,00
160,00
180,00

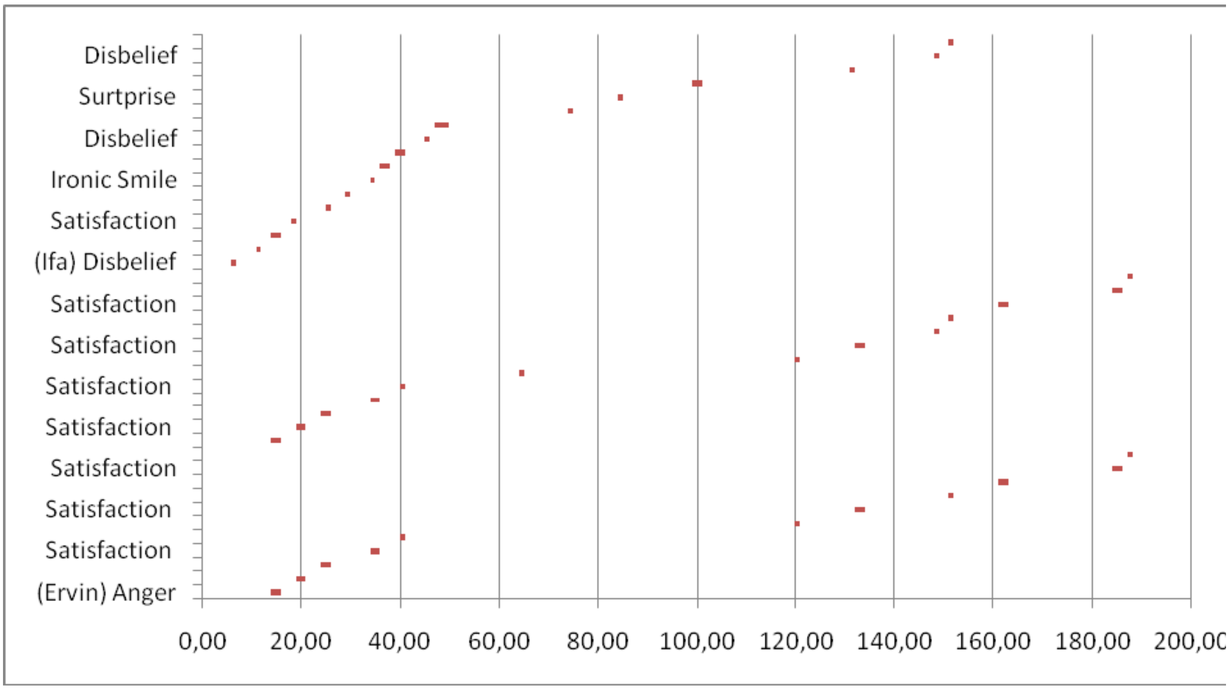
Anna 6


(Ifa) Disgust
Astonishment
Disbelief
Surprise
Satisfaction
Happiness
(Ifa) Sadness
Astoishment
Satisfaction
Satisfaction
Ironic Smile
(Ervine) Disbelief
Disbelief
Astonishment
Satisfaction
(Ervin) Sadness
0,00
20,00
40,00
60,00
80,00
100,00
120,00

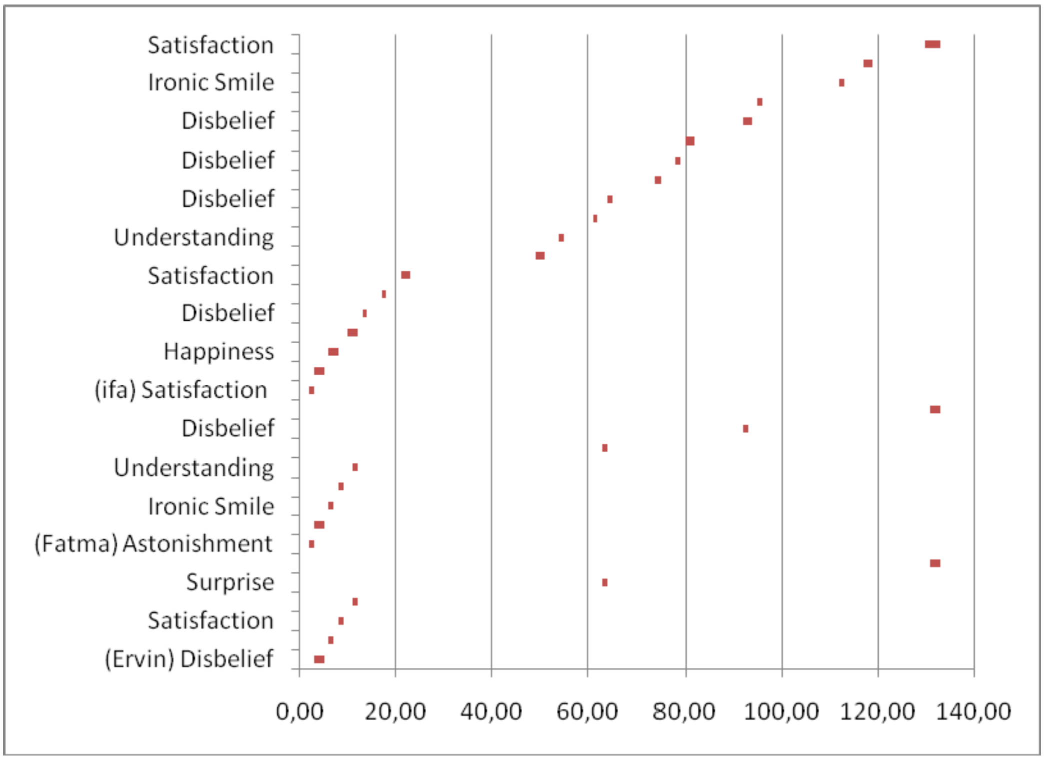
Anna 7


(Ifa) Disbelief
Disbelief
Satisfaction
Disbelief
Disbelief
Astonishment
Astonishment
(Ifa) Satisfaction
Ironic Smile
Satisfaction
Satisfaction
Diebelief
(Ervine) Understanding
Satisfaction
Understanding
(Ervin) Disbelief
0,00
20,00
40,00
60,00
80,00
100,00
120,00
140,00

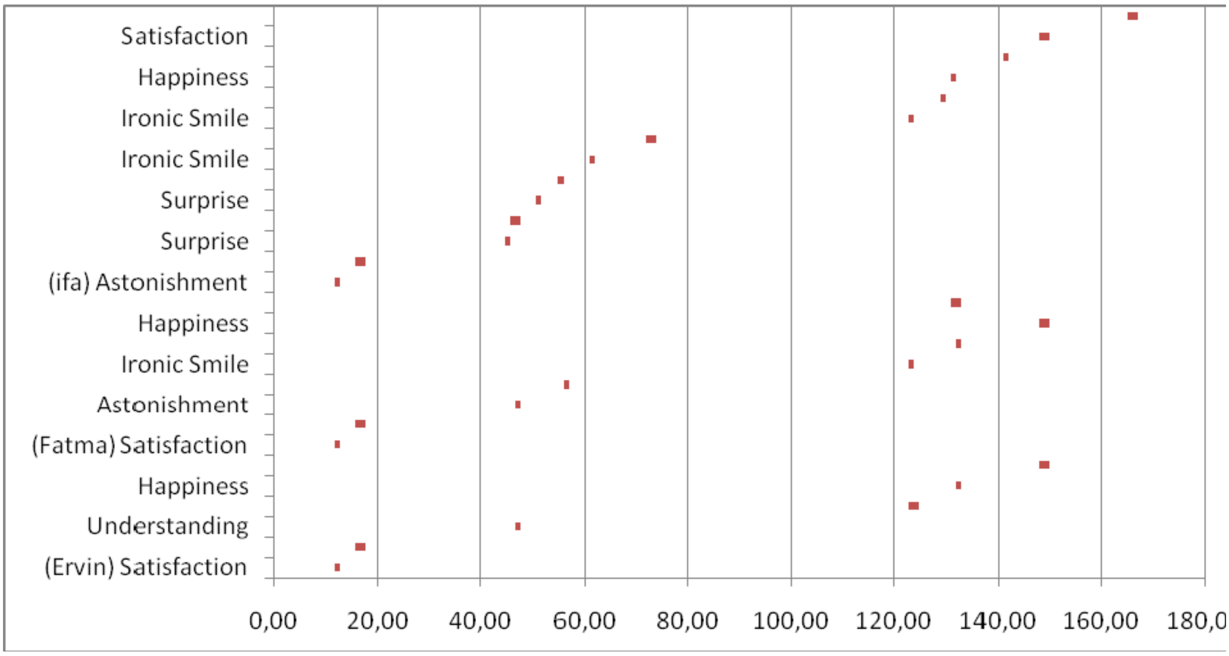
Anna 8


Disbelief
Surtprise
Disbelief
Ironic Smile
Satisfaction
(Ifa) Disbelief
Satisfaction
Satisfaction
Satisfaction
Satisfaction
Satisfaction
Satisfaction
Satisfaction
(Ervin) Anger
0,00
20,00
40,00
60,00
80,00
100,00
120,00
140,00
160,00
180,00
200,0C

Anna 9



Anna 10


Satisfaction
Happiness
Ironic Smile
Ironic Smile
Surprise
Surprise
(ifa) Astonishment
Happiness
Ironic Smile
Astonishment
(Fatma) Satisfaction
Happiness
Understanding
(Ervin) Satisfaction